

Density-gradient theory: a macroscopic approach to quantum confinement and tunneling in semiconductor devices

M.G. Ancona

Published online: 23 April 2011
© Springer Science+Business Media LLC 2011

Abstract Density-gradient theory provides a macroscopic approach to modeling quantum transport that is particularly well adapted to semiconductor device analysis and engineering. After some introductory observations, the basis of the theory in macroscopic and microscopic physics is summarized, and its scattering-dominated and scattering-free versions are introduced. Remarks are also given about the underlying mathematics and numerics. A variety of applications of the theory to both quantum confinement and quantum tunneling situations are then reviewed. In doing so, particular emphasis is put on understanding the range of validity of the theory and on its unexpected power as a phenomenology. The article closes with a few comments about the future.

Keywords Continuum · Density-gradient · Electron transport · Semiconductor device simulation · Quantum confinement · Quantum tunneling · Thermodynamics

1 Introduction

1.1 Electron transport modeling

To understand, design, and optimize electronic devices and circuits it is essential that one have a quantitative description of how electrons move in semiconductors and in their connecting wires and insulators. The subject of this review is one such description known as **density-gradient (DG) theory** that is particularly well suited to the engineering of semiconductor devices that are small enough to be impacted

by the direct effects of quantum mechanics, and especially by the phenomena of quantum confinement and quantum tunneling.

The various mathematical descriptions of electron flow in biased semiconductors that have been proposed and applied over the past half-century can be usefully divided into two types: *theories* and *phenomenologies*. Theories are distinguished by the fact that their equations are grounded in physical principles and can thereby be predictive. By contrast, phenomenologies use mathematics (often drawn willy-nilly from well-founded theories) merely as fitting functions for regressing physical data and are thus solely of interpolative value. The DG approach can be used in both of these ways, but it is best understood by stressing its foundations and its status as a theory.

As in other areas of mathematical physics, the theories of electron transport can be further split according to whether they are *microscopic* or *macroscopic* in character (see Table 1). The key distinction is in the nature of the primitive elements that form the theory; microscopic theories deal with the *individual* electrons (or electron wave functions, density matrices, etc.) whereas their macroscopic counterparts are framed in terms of electron *populations*. An additional important sub-division among microscopic theories is that into classical and quantum theories based on whether or not it is possible to localize the individual electrons in phase space. Some prime examples are listed in Table 1. That macroscopic theories have no discrete electrons means a similar scission among these theories makes no sense. Instead a bifurcation exists into lumped and continuum theories based on the *size* of the electron populations, with DG theory falling into the continuum category as indicated in Table 1. An obvious key requirement for the existence of a macroscopic electron transport theory is that there be enough electrons to form a population(s) with meaningful average prop-

M.G. Ancona (✉)
Naval Research Laboratory, Washington, DC 20375, USA
e-mail: ancona@estd.nrl.navy.mil

| Report Documentation Page | | | Form Approved OMB No. 0704-0188 | | |
|--|------------------------------------|--|---|---|---------------------------------|
| Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. | | | | | |
| 1. REPORT DATE APR 2011 | | 2. REPORT TYPE | | 3. DATES COVERED 00-00-2011 to 00-00-2011 | |
| 4. TITLE AND SUBTITLE Density-gradient theory: a macroscopic approach to quantum confinement and tunneling in semiconductor devices | | 5a. CONTRACT NUMBER | | | |
| | | 5b. GRANT NUMBER | | | |
| | | 5c. PROGRAM ELEMENT NUMBER | | | |
| 6. AUTHOR(S) | | 5d. PROJECT NUMBER | | | |
| | | 5e. TASK NUMBER | | | |
| | | 5f. WORK UNIT NUMBER | | | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Research Laboratory, Washington, DC, 20375 | | 8. PERFORMING ORGANIZATION REPORT NUMBER | | | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | | 10. SPONSOR/MONITOR'S ACRONYM(S) | | | |
| | | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) | | | |
| 12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited | | | | | |
| 13. SUPPLEMENTARY NOTES | | | | | |
| 14. ABSTRACT Density-gradient theory provides a macroscopic approach to modeling quantum transport that is particularly well adapted to semiconductor device analysis and engineering. After some introductory observations, the basis of the theory in macroscopic and microscopic physics is summarized, and its scattering-dominated and scatteringfree versions are introduced. Remarks are also given about the underlying mathematics and numerics. A variety of applications of the theory to both quantum confinement and quantum tunneling situations are then reviewed. In doing so particular emphasis is put on understanding the range of validity of the theory and on its unexpected power as a phenomenology. The article closes with a few comments about the future. | | | | | |
| 15. SUBJECT TERMS Continuum, Density-gradient, Electron, transport, Semiconductor device simulation, Quantum confinement, Quantum tunneling, Thermodynamics | | | | | |
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT Same as Report (SAR) | 18. NUMBER OF PAGES 33 | 19a. NAME OF RESPONSIBLE PERSON |
| a. REPORT unclassified | b. ABSTRACT unclassified | c. THIS PAGE unclassified | | | |

Table 1 Classification of electron transport theories

| Macroscopic | | Microscopic | |
|---|-----------------------------------|---|--|
| Lumped | Continuum | Classical | Quantum |
| Equivalent circuits, transmission lines | Diffusion-drift, density-gradient | Semi-classical electron dynamics, Boltzmann transport | Schrödinger, density-matrix, Wigner function, NEGF |

erties. This foundational consideration—known in the context of continuum theories as the *continuum assumption*—has many subtleties that are beyond the scope of this paper. We note only that DG theory often pushes its continuum assumption well beyond where one might expect it to fail, and this is possible largely because of the small mass of the electron and the consequent strong “smearing” effect of quantum mechanics. In any event, our approach with the continuum assumption is simply to assert its validity, and to look for justification only *a posteriori* (if at all) based on the theory’s predictions and results. Of course there are certain situations (e.g., molecular conduction) for which a continuum approach will be patently inappropriate.

1.2 Quantum transport

The three main “quantum” behaviors of an electron gas in a semiconductor—all of course well known—that one would like to describe with a quantum transport theory are:

- Quantum compressibility.* Quantum mechanics induces an electronic repulsion (via the Pauli principle) that makes electron gases in solids harder to compress than an equally dense thermal distribution of classical point particles.
- Electron evanescence.* Evanescence is a facet of the wave nature of electrons that, as in the analogous phenomena in acoustics and optics, arises when electron waves encounter a “barrier” region incapable of sustaining their propagation. Macroscopically this effect is commonly manifested as the phenomena of quantum confinement and quantum tunneling.
- Electron diffraction/interference.* The wave nature of electrons can also give rise to diffraction and interference effects. From a macroscopic standpoint, the primary manifestation is in the effective mass, with other macroscopic consequences being seen only in the rare circumstance of a coherent electron gas.

Of these effects, quantum compressibility is the easiest to incorporate in a macroscopic description, and is in fact ordinarily included in diffusion-drift theory simply by using a Fermi-Dirac equation of state. Encompassing the evanescent phenomena of confinement and tunneling within a macroscopic theory are the prime purposes of DG theory, and thus these constitute the main subject matter of this review.

Lastly, diffraction/interference effects are least amenable to a macroscopic description with only the effective mass aspect being easily incorporated. In any event, in all such usages of macroscopic theory it is important to emphasize that the goal is emphatically not to replicate or replace quantum mechanics, but rather to provide a convenient and physically well-founded means for describing the quantum phenomena seen in electronic devices.

Since the electrons that form the populations we are concerned with are incoherent, the only possible physical principles on which a macroscopic electron transport description can be based are *classical* ones—primarily the conservation laws of mass, momentum and energy. For this reason, macroscopic-continuum theories like DG theory are often referred to as classical field theories [1]. We are thus led to the seemingly contradictory statement that *DG theory is a classical theory of quantum phenomena*. That such a thing is possible is perhaps best illustrated by considering the example of barrier tunneling. Of course, microscopically this process cannot be treated classically, e.g., it is well known that a tunneling electron experiences a momentary violation of energy conservation.¹ But as already noted, a macroscopic theory is concerned not with individual electrons but with electron populations for which any such violation might well be negligible. For instance, in most tunneling devices the time scales of interest are set by macroscopic RC delays that are far longer than the microscopic tunneling time. Hence a description of the device’s physics would not be materially affected if the tunneling were regarded as instantaneous, in which case the electron gas would always conserve energy as the classical laws demand. Thus, in general, the question is whether the violations of conservation laws that characterize the quantum behavior at the particle level average out and can thereby be neglected at the population level. In these terms the issue is much like the theory’s other continuum assumptions that we assert to hold at least for some important device situations.

1.3 Density-gradient approach

Diffusion-drift (DD) theory is the original and most common continuum theory of electron transport in semicon-

¹This violation of course disappears in quantum field theory when one includes virtual particles in the description.

ductors. It originated in work by Schottky, and was formalized by Shockley and Van Roosbroeck in the early 1950s [2, 3]. As will be discussed in Sect. 2.1, in its simplest form DG theory represents a direct generalization of DD theory in which the equation of state of the electron (or hole) gas depends not only on its density but also on its density-gradient. This gradient dependence introduces non-locality, and in this way can be used to represent quantum non-locality to lowest order. With this simple change, the equations are found to exhibit new behaviors, including ones with the physical characteristics of quantum confinement and tunneling. Mathematically these new solution behaviors are simply the consequence of the introduced gradients raising the order of the differential system and thereby inducing “quantum” boundary layers.

DG theory’s central concept of using gradients as a lowest-order representation of physical non-locality is a powerful idea with many antecedents. The basic notion was appreciated first by Maxwell in his pioneering work on the kinetic theory of gases wherein he showed that the equation of state of even a monatomic hard-sphere gas has gradient corrections [4]. In the years since this fruitful idea has recurred in many areas of mathematical physics in both classical and quantum mechanical contexts. Of especial interest for us is an approach to quantum mechanics dating from the 1930s in the work of Wigner [5], Bloch [6] and Weizacker [7] who, faced with the difficulties of solving the Schrödinger equation, devised approximations to it based on gradient expansions. Although one goal of such work (beginning with Weizacker) was in fact macroscopic—namely, to derive density-gradient corrections to the electron gas equation of state—that there existed an associated macroscopic transport theory was not appreciated until the late 1980s with the development of DG theory [8].²

The subject of this review is DG theory as an engineering tool for modeling electronic devices. The paper is organized as follows. In Sect. 2, the basic equations of DG theory are summarized in general form with emphasis on the important distinction of classical field theory between physical principles and material response functions. The following two main sections are devoted to applications of DG theory, with the first (Sect. 3) covering quantum confinement situations and the second (Sect. 4) treating quantum tunneling situations. The paper concludes in Sect. 5 with a few remarks including about future research directions.

²It is perhaps worth noting that the use of gradient expansions in microscopic work to represent the kinetic energy operator of quantum mechanics faded in the 1960s with the advent of density-functional theory [9], and especially the Kohn-Sham formalism. Interestingly, the latter also led to fresh uses for gradient expansions as a means of approximating the non-local exchange-correlation functional, and in this form they remain a fixture of electronic structure calculations [10].

2 Density-gradient theory

DG theory is an example of a classical field theory [1], and as such it is formed of three basic parts: Primitive elements, physical principles, and material response functions. The primitive elements are the mathematical field variables (e.g., densities, forces, etc.) that are postulated to quantify the constituents and interactions of the model. As their name implies, the physical principles (Sect. 2.1) are the constraints imposed on the primitive elements by the laws of physics, and it is the breadth of applicability of these laws that endows the overall theory with predictive power. The final and most subtle part of a classical field theory is the material response functions (Sect. 2.2). In principle, these functions are fairly arbitrary and are constrained only by considerations of consistency, symmetry and invariance. The weakness of the latter constraints can allow a considerable amount of curve-fitting to creep in, and therefore there is a crucial additional bias toward *simple* material response functions. Most commonly, this simplicity is embodied in functions that are *linear* and *instantaneous*. Broadly speaking, a “good” classical field theory is one whose primitive elements and physical principles are such that simple material response functions lead to an accurate and widely applicable description.

DG theory is exactly like DD theory in defining a semiconductor as consisting of three interpenetrating continua: An electron gas, a hole gas, and a rigid lattice continuum.³ These constituents are assumed characterized by various densities (e.g., of charge, momentum or energy), and as noted earlier this is a major assertion of the theory (known as continuum assumption). Moreover, the constituents are assumed to interact through various forces (e.g., between neighboring elements of the electron gas) and sources/sinks (e.g., mediating the transfer of charge between constituents) with the key generalization distinguishing DG theory from DD theory being the form of the force interactions assumed to exist within the carrier gases.

2.1 Physical principles

As emphasized in Sect. 1.2, the physical principles of a macroscopic description of quantum transport are necessarily classical; in particular, they consist of the conservation of charge/mass and momentum plus the laws of electrostatics and thermodynamics. These physical principles are expressed mathematically in terms of the primitives of the theory, and the most general form of this mathematics is as integral equations [1, 8]. With one exception (see below), the physical principles of DG theory are as follows.

³Both DD and DG theories are readily generalized to include multiple electron and hole gases, e.g., see [11].

Charge/mass balance. The equations expressing the conservation of charge/mass for the electron and hole gases readily reduce to the following expressions:

$$\begin{aligned} \frac{\partial}{\partial t} \int_V n dV &= - \int_S \mathbf{n} \cdot n \mathbf{v}_n dS + \int_V G_{npl} dV \quad \text{and} \\ \frac{\partial}{\partial t} \int_V p dV &= - \int_S \mathbf{n} \cdot p \mathbf{v}_p dS + \int_V G_{npl} dV \end{aligned} \quad (2.1.1)$$

where n and p are the number densities and \mathbf{v}_n and \mathbf{v}_p are the average velocities of the electron and hole gases, respectively, and \mathbf{n} is the outward unit normal to the surface S . In words, these equations say that the total electron or hole charge in the arbitrary volume V increases in time due to inflow through its surface S , and via generation processes inside V . The lattice's impurities are assumed fully ionized so that the generation term (G_{npl}) includes only the net pair generation rate. Also, since the lattice is assumed immobile, the lattice charge/mass balance equation is trivially satisfied.

Linear momentum balance. The momentum balance equations for the electron and hole gases are⁴

$$\begin{aligned} \frac{\partial}{\partial t} \int_V m_n n \mathbf{v}_n dV &= - \int_S \mathbf{n} \cdot m_n n \mathbf{v}_n \mathbf{v}_n dS + \int_S \mathbf{n} \cdot \tau^n dV \\ &\quad - \int_V qn \mathbf{E} dV + \int_V qn \mathbf{E}_n dV \\ \frac{\partial}{\partial t} \int_V m_p p \mathbf{v}_p dV &= - \int_S \mathbf{n} \cdot m_p p \mathbf{v}_p \mathbf{v}_p dS + \int_S \mathbf{n} \cdot \tau^p dV \\ &\quad + \int_V qp \mathbf{E} dV + \int_V qp \mathbf{E}_p dV \end{aligned} \quad (2.1.2)$$

where τ^n and τ^p are the stress tensors, $-qn\mathbf{E}$ and $qp\mathbf{E}$ are the forces exerted by the electrostatic field, and $qn\mathbf{E}_n$ and $qp\mathbf{E}_p$ are drag forces that are the macroscopic resultants of innumerable microscopic scattering events that act to impede the flow of carriers through the lattice. In words, these equations are expressions of Newton's Second Law that hold that the time rates of increase of the gas momenta in the volume V are equal to the net influxes of momentum through its surface S plus the supplies of momentum to the gases from the forces exerted by the electron gas, the electrostatic field, and the lattice. With the lattice assumed immobile, its momentum balance equation is trivially satisfied.

Angular momentum balance. In the interest of brevity we omit the integral forms expressing angular momentum conservation in the electron and hole gases. As will be

seen, their only consequence (when magnetic effects are not present) is a demand that the stress tensors be symmetric.

Electrostatics. The familiar integral conditions expressing Gauss's Law of electrostatics and Faraday's law in the electrostatic limit are:

$$\begin{aligned} \int_S \mathbf{n} \cdot \mathbf{D} dS &= \int_V q(N - n + p) dV \quad \text{and} \\ \oint_C \mathbf{E} \cdot d\mathbf{s} &= 0 \end{aligned} \quad (2.1.3)$$

where N is the charge density of the lattice (usually due mostly to ionized impurities) and $\mathbf{D} = \mathbf{E} + \mathbf{P}$.

Energy balance. In this paper we assume the temperature to be uniform so that all macroscopic energy is mechanical (i.e., heat energy need not be considered explicitly) and the energy balance equations do not provide additional dynamical information. Nevertheless, they are needed for understanding the form of the material response functions, and are especially important when the electron and hole gas interactions include the double-pressures that are mechanically self-equilibrating (and so do not appear in (2.1.2)) yet are still capable of storing energy. The integral conditions expressing the conservation of energy for the electron and hole gases in DG theory are:

$$\begin{aligned} \frac{\partial}{\partial t} \int_V \left(n\varepsilon_n + \frac{1}{2} m_n n \mathbf{v}_n \cdot \mathbf{v}_n \right) dV &= - \int_S \mathbf{n} \cdot \mathbf{v}_n \left(n\varepsilon_n + \frac{1}{2} m_n n \mathbf{v}_n \cdot \mathbf{v}_n \right) dS \\ &\quad + \int_S \mathbf{n} \cdot (\tau^n \cdot \mathbf{v}_n - \eta^n \nabla \cdot \mathbf{v}_n) dS \\ &\quad - \int_V qn \mathbf{v}_n \cdot \mathbf{E} dV + \int_V H_{nl} dV \\ \frac{\partial}{\partial t} \int_V \left(p\varepsilon_p + \frac{1}{2} m_p p \mathbf{v}_p \cdot \mathbf{v}_p \right) dV &= - \int_S \mathbf{n} \cdot \mathbf{v}_p \left(p\varepsilon_p + \frac{1}{2} m_p p \mathbf{v}_p \cdot \mathbf{v}_p \right) dS \\ &\quad + \int_S \mathbf{n} \cdot (\tau^p \cdot \mathbf{v}_p - \eta^p \nabla \cdot \mathbf{v}_p) dS \\ &\quad + \int_V qp \mathbf{v}_p \cdot \mathbf{E} dV + \int_V H_{pl} dV \end{aligned} \quad (2.1.4)$$

These equations say that the time rate of change of total energy (internal + kinetic) in V is equal to the net rate of energy flow through S plus the rates of working of the various

⁴For simplicity we have omitted treatment of the effective mass. Its proper macroscopic origin is in the electron-lattice interaction (expressing the effect of electron diffraction by the lattice) of which we have included explicitly only the dissipative portion as the last term in (2.1.2). See [12] for further discussion.

forces and the double-pressure vectors η^n and η^p .⁵ Lastly for the lattice (which again is assumed not to move) we have

$$\begin{aligned} \frac{\partial}{\partial t} \int_V \rho \varepsilon_l dV &= \int_V \mathbf{E} \cdot \frac{\partial \mathbf{P}}{\partial t} dS - \int_V (H_{nl} + H_{pl}) dV \\ &+ \int_V H_{npl} dV \end{aligned} \quad (2.1.5)$$

Second law of thermodynamics. Because the second law of thermodynamics is a “negative” constraint about what should *not* happen, it is best to defer its treatment until after the material response functions are discussed.

2.2 Material response functions

The equations of DG theory derived in the previous section form a mathematically indeterminate system, i.e., with more variables than equations. This indeterminacy reflects the fact that the general physical principles do not dictate the state of the system, nor should they for otherwise every semiconductor would be identical. To supply the material-specific information, and thereby complete the system, one must adjoin auxiliary equations known as material response functions. These extra equations are arbitrary apart from certain restrictions of consistency, symmetry and invariance. Of these constraints, the most important for us is the demand that the material response functions be thermodynamically admissible.

The first step toward understanding the thermodynamic constraints on the material response functions is to write down the energy balance equation for the entire system. To this end, we add up the differential versions of (2.1.4)–(2.1.5), substitute (2.1.1) appropriately, and split off isotropic pressures using the definitions $\tau^n \equiv -P_n \mathbf{I} + \sigma^n$ and $\tau^p \equiv -P_p \mathbf{I} + \sigma^p$. Using indicial notation for clarity, the resulting total local energy balance equation is

$$\begin{aligned} \rho \frac{d^l \varepsilon_l}{dt} + n \frac{d^n \varepsilon_n}{dt} + p \frac{d^p \varepsilon_p}{dt} - \frac{P_n}{n} \frac{d^n n}{dt} - \frac{P_p}{p} \frac{d^p p}{dt} \\ - \frac{\eta_{j,j}^n}{n} \frac{d^n n_{,j}}{dt} - \frac{\eta_{j,j}^p}{p} \frac{d^p p_{,j}}{dt} - E_j \frac{d^l P_j}{dt} \\ - v_{n,j,i} \left[\sigma_{ij}^n - n \left(\frac{\eta_k^n}{n} \right)_{,k} \delta_{ij} + \frac{\eta_i^n n_{,j}}{n} \right] \\ - v_{p,j,i} \left[\sigma_{ij}^p - p \left(\frac{\eta_k^p}{p} \right)_{,k} \delta_{ij} + \frac{\eta_i^p p_{,j}}{p} \right] \\ = -qnE_{n,j} v_{n,j} - qpE_{p,j} v_{p,j} \end{aligned}$$

⁵The forms of the terms in (2.1.4) that dictate how the higher-order stresses contribute to the energy balance is best understood from a variational argument given in [8]. Alternatively, it can be regarded as justified by the results that (2.1.4) leads to in Sect. 2.3.

$$\begin{aligned} - G_{npl,j} \left(\frac{\eta_j^n}{n} + \frac{\eta_j^p}{p} \right) + H_{npl} - G_{npl} \left(\varepsilon_n + \frac{P_n}{n} \right. \\ \left. - \frac{1}{2} m_n v_{n,j} v_{n,j} + \varepsilon_p + \frac{P_p}{p} - \frac{1}{2} m_p v_{p,j} v_{p,j} \right) \end{aligned} \quad (2.2.1)$$

Since the stresses and double-pressures act in a purely non-dissipative manner (e.g., assuming there are no viscous effects) (2.2.1) divides cleanly into the terms on the left side of the equality that are purely recoverable, and those on the right side that are purely dissipative. The recoverability of the terms on the left side implies path independence and integrability, and hence the existence of an entropy function η with

$$\begin{aligned} \rho \frac{d^l \varepsilon_l}{dt} + n \frac{d^n \varepsilon_n}{dt} + p \frac{d^p \varepsilon_p}{dt} - \frac{P_n}{n} \frac{d^n n}{dt} - \frac{P_p}{p} \frac{d^p p}{dt} \\ - \frac{\eta_{j,j}^n}{n} \frac{d^n n_{,j}}{dt} - \frac{\eta_{j,j}^p}{p} \frac{d^p p_{,j}}{dt} - E_j \frac{d^l P_j}{dt} \\ - v_{n,j,i} \left[\sigma_{ij}^n - n \left(\frac{\eta_k^n}{n} \right)_{,k} \delta_{ij} + \frac{\eta_i^n n_{,j}}{n} \right] \\ - v_{p,j,i} \left[\sigma_{ij}^p - p \left(\frac{\eta_k^p}{p} \right)_{,k} \delta_{ij} + \frac{\eta_i^p p_{,j}}{p} \right] \\ = \rho T \frac{d\eta}{dt} \end{aligned} \quad (2.2.2)$$

where the temperature T is the integrating factor. Combining (2.2.1) and (2.2.2) we further obtain an expression for the rate of entropy production

$$\begin{aligned} \rho \frac{d\eta}{dt} = -\frac{1}{T} \left[qnE_{n,j} v_{n,j} + qpE_{p,j} v_{p,j} + G_{npl,j} \left(\frac{\eta_j^n}{n} + \frac{\eta_j^p}{p} \right) \right. \\ \left. - H_{npl} + G_{npl} \left(\varepsilon_n + \frac{P_n}{n} - \frac{1}{2} m_n v_{n,j} v_{n,j} + \varepsilon_p \right. \right. \\ \left. \left. + \frac{P_p}{p} - \frac{1}{2} m_p v_{p,j} v_{p,j} \right) \right] \geq 0 \end{aligned} \quad (2.2.3)$$

where the inequality—often called the Clausius-Duhem inequality—is the local version of the thermodynamic prescription on decreasing entropy.

When the terms in the energy balance equation are as above being either purely recoverable or purely dissipative, the constitutive equations split along similar lines with the recoverable constitutive equations being constrained by (2.2.2) and the dissipative constitutive equations by (2.2.3) as discussed next.

Recoverable constitutive equations. In order for σ^n and σ^p to depend only on the density-gradients as desired (and not on the individual components of the strain-gradients),

the terms in (2.2.2) multiplying the velocity-gradients must vanish which implies:

$$\begin{aligned}\tau_{ij}^n &= \left[-P_n + n \left(\frac{\Xi_n n_{,k}}{n} \right)_{,k} \right] \delta_{ij} - \frac{\Xi_n}{n} n_{,i} n_{,j} \\ \tau_{ij}^p &= \left[-P_p + p \left(\frac{\Xi_p p_{,k}}{p} \right)_{,k} \right] \delta_{ij} - \frac{\Xi_p}{p} p_{,i} p_{,j}\end{aligned}\quad (2.2.4)$$

where we have used the previously noted symmetry of the stress tensors (from angular momentum balance) to write $\eta_k^n \equiv \Xi_n n_{,k}$ and $\eta_k^p \equiv \Xi_p p_{,k}$. Defining the thermodynamic state function $\chi_l \equiv \varepsilon_l - \mathbf{P}_j \cdot \mathbf{E}_j - \eta T$, (2.2.2) with (2.2.4) implies:

$$\begin{aligned}\rho \frac{d^l \chi_l}{dt} + n \frac{d^n \varepsilon_n}{dt} + p \frac{d^p \varepsilon_p}{dt} - \frac{P_n}{n} \frac{d^n n}{dt} - \frac{P_p}{p} \frac{d^p p}{dt} \\ - \frac{\Xi_n}{2n} \frac{d^n \Pi_n}{dt} - \frac{\Xi_p}{2p} \frac{d^p \Pi_p}{dt} + \mathbf{P}_j \cdot \frac{d^l \mathbf{E}_j}{dt} + \rho \eta \frac{dT}{dt} = 0\end{aligned}\quad (2.2.5)$$

where $\Pi_n \equiv n_{,i} n_{,i}$ and $\Pi_p \equiv p_{,i} p_{,i}$. The form of (2.2.5) then indicates that the lattice acts as a simple dielectric with $\chi_l = \chi_l(\mathbf{E}, T)$, and that the electron and hole gases are well described by:

$$\varepsilon_n = \varepsilon_n(n, \Pi_n, T), \quad \varepsilon_p = \varepsilon_p(p, \Pi_p, T) \quad (2.2.6)$$

Using the chain rule, (2.2.5) can then be re-expressed as

$$\begin{aligned}\left(n \frac{\partial \varepsilon_n}{\partial n} - \frac{P_n}{n} \right) \frac{d^n n}{dt} + \left(p \frac{\partial \varepsilon_p}{\partial p} - \frac{P_p}{p} \right) \frac{d^p p}{dt} \\ + \left(n \frac{\partial \varepsilon_n}{\partial \Pi_n} - \frac{\Xi_n}{2n} \right) \frac{d^n \Pi_n}{dt} + \left(p \frac{\partial \varepsilon_p}{\partial \Pi_p} - \frac{\Xi_p}{2p} \right) \frac{d^p \Pi_p}{dt} \\ + \left(\rho \frac{\partial \chi_l}{\partial \mathbf{E}} + \mathbf{P}_j \right) \cdot \frac{d^l \mathbf{E}_j}{dt} + \left(\rho \eta + \rho \frac{\partial \chi_l}{\partial T} + n \frac{\partial \varepsilon_n}{\partial T} \right. \\ \left. + p \frac{\partial \varepsilon_p}{\partial T} \right) \frac{dT}{dt} = 0\end{aligned}\quad (2.2.7)$$

For this equation to hold under all circumstances it must be that the coefficients of the time derivatives vanish, and the recoverable constitutive equations then follow:

$$\begin{aligned}P_n &= n^2 \frac{\partial \varepsilon_n}{\partial n}, \quad P_p = p^2 \frac{\partial \varepsilon_p}{\partial p} \\ \Xi_n &= 2n^2 \frac{\partial \varepsilon_n}{\partial \Pi_n}, \quad \Xi_p = 2p^2 \frac{\partial \varepsilon_p}{\partial \Pi_p} \\ \mathbf{P} &= -\frac{\partial \chi_l}{\partial \mathbf{E}}, \quad \rho \eta = -\rho \frac{\partial \chi_l}{\partial T} - n \frac{\partial \varepsilon_n}{\partial T} - p \frac{\partial \varepsilon_p}{\partial T}\end{aligned}\quad (2.2.8)$$

Dissipative constitutive equations. The rate of entropy production inequality (2.2.3) can be re-written as:

$$\begin{aligned}\rho \frac{\partial \eta}{\partial t} &= -\frac{1}{T} \left[qn \mathbf{E}_{n_j} v_{n_j} + qp \mathbf{E}_{p_j} v_{p_j} \right. \\ &\quad + G_{npl,j} \left(\frac{\Xi_n n_{,j}}{n} + \frac{\Xi_p p_{,j}}{p} \right) - H_{npl} \\ &\quad + G_{npl} \left(\varepsilon_n + \frac{P_n}{n} - \frac{1}{2} m_n v_{n_j} v_{n_j} + \varepsilon_p + \frac{P_p}{p} \right. \\ &\quad \left. \left. - \frac{1}{2} m_p v_{p_j} v_{p_j} \right) \right] \geq 0\end{aligned}\quad (2.2.9)$$

The form of this equation indicates that the primary dependences of the dissipative constitutive equations are

$$\begin{aligned}\mathbf{E}_n &= \mathbf{E}_n(\mathbf{v}_n), \quad \mathbf{E}_p = \mathbf{E}_p(\mathbf{v}_p) \\ G_{npl} &= G_{npl}(n, \Pi_n, p, \Pi_p)\end{aligned}\quad (2.2.10)$$

however these forms do not preclude possible dependences on other variables, and any such dependences are permitted so long as the inequality (2.2.9) is satisfied. Insisting that each term in (2.2.9) is individually positive ensures the total is also, and yields

$$\begin{aligned}\mathbf{E}_n(\mathbf{v}_n) \cdot \mathbf{v}_n &\leq 0, \quad \mathbf{E}_p(\mathbf{v}_p) \cdot \mathbf{v}_p \leq 0 \\ H_{npl} - \nabla G_{npl} \left(\frac{\Xi_n \nabla n}{n} + \frac{\Xi_p \nabla p}{p} \right) - G_{npl} \left(\varepsilon_n + \frac{P_n}{n} \right. \\ &\quad \left. - \frac{1}{2} m_n v_{n_j} v_{n_j} + \varepsilon_p + \frac{P_p}{p} - \frac{1}{2} m_p v_{p_j} v_{p_j} \right) \geq 0\end{aligned}\quad (2.2.11)$$

Examples. To illustrate the material response functions with some specific examples, we first note that since DG theory must reduce to DD theory when the density-gradients are small (and inertia is negligible), a basic guideline in developing specific material response functions for DG theory is to use those of DD theory but possibly include density-gradient corrections. And as we shall see, this simple approach spawns theories that are often quite accurate.

- (i) Equations of state. The most important material response functions of DG theory are the “equations of state” that characterize the electron and hole gases. As we have seen, these equations are no longer DD theory’s simple relationships between pressure and density, but instead generalize into expressions relating stress to the density and its spatial derivatives. Moreover, we found that the dependence on the spatial derivatives was not arbitrary, but had to be formed of specific combinations of derivatives of the internal energies as given in (2.2.8) with (2.2.6). For these energy functions to reduce to those of DD theory we assume without loss of generality that

$$\begin{aligned}\varepsilon_n(n, \Pi_n) &= \varepsilon_n^{DD}(n) + \varepsilon_n^{DG}(n, \Pi_n) \quad \text{and} \\ \varepsilon_p(p, \Pi_p) &= \varepsilon_p^{DD}(p) + \varepsilon_p^{DG}(p, \Pi_p)\end{aligned}\quad (2.2.12)$$

where $\varepsilon_n^{DD}(n)$ and $\varepsilon_p^{DD}(p)$ are the gradient-independent functions of DD theory, and $\varepsilon_n^{DG}(n, \Pi_n)$ and $\varepsilon_p^{DG}(p, \Pi_p)$ are the correction terms that vanish as the density-gradients go to zero and we have suppressed the temperature dependences for clarity. The simplest form for these energies are ones that yield linear relationships between pressure and density, and between double-pressure and density-gradient; that the former defines an ideal gas leads us to refer to the latter as defining an **ideal gradient gas**:

$$P_n = k_B T n \quad \text{and} \quad P_p = k_B T p \quad (\text{ideal gases}) \quad (2.2.13a)$$

$$\eta^n = b_n \nabla n \quad \text{and} \quad \eta^p = b_p \nabla p \quad (\text{ideal gradient gases}) \quad (2.2.13b)$$

where b_n and b_p are (linear) **density-gradient (DG)** coefficients that characterize the strength of the gradient responses of the gases. In general, these latter coefficients could be second-rank tensors and, as we shall see in Sect. 3.4, this possibility can be helpful when the DG equations are used phenomenologically. Equations (2.2.13a) and (2.2.13b) imply the energy functions

$$\varepsilon_n^{DD}(n) = k_B T \ln\left(\frac{n}{n_0}\right) \quad \text{and} \quad (2.2.14a)$$

$$\varepsilon_p^{DD}(p) = k_B T \ln\left(\frac{p}{p_0}\right)$$

$$\varepsilon_n^{DG}(n, \Pi_n) = \frac{b_n}{2} \frac{\Pi_n}{n^2} \quad \text{and} \quad (2.2.14b)$$

$$\varepsilon_p^{DG}(p, \Pi_p) = \frac{b_p}{2} \frac{\Pi_p}{p^2}$$

where n_0 and p_0 are constants. With higher densities and density-gradients, the linear theories are no longer such good approximations, and modifications to (2.2.13a), (2.2.13b), (2.2.14a) and (2.2.14b) must be considered. The familiar example is the corrections that enter for high density due to Fermi-Dirac statistics. Very little comparable work has been done on nonlinear DG theories.

- (ii) Polarization. In the electrostatic limit the dielectric properties are almost always well approximated by the usual linear, instantaneous relation

$$\mathbf{P} = \chi_d \mathbf{E} \quad \text{and} \quad \mathbf{D} = \varepsilon_d \mathbf{E} \quad (2.2.15)$$

where χ_d is the electric susceptibility and $\varepsilon_d \equiv \varepsilon_0 + \chi_d$ is the electric permittivity.

- (iii) Drag forces. As in DD theory [13], the simplest drag expressions are the linear, instantaneous forms:

$$\mathbf{E}_n = -\mathbf{v}_n / \mu_n \quad \text{and} \quad \mathbf{E}_p = -\mathbf{v}_p / \mu_p \quad (2.2.16)$$

where μ_n and μ_p must be positive by virtue of the entropy inequalities in (2.2.10) and are often taken to depend on the electric field, e.g., to represent velocity saturation. Whether such mobility models are modified in value or form in DG theory is as yet unexplored.

- (iv) Generation-recombination. In DD theory the forms for G_{npl} are typically nonlinear and sometimes non-instantaneous [13]. Most such models can be carried over into DG theory with a simple proviso that we illustrate using the fairly general expression

$$G_{npl} = g_{npl}[n_{eq} p_{eq} - np] \quad (2.2.17)$$

where the quantity g_{npl} depends on the particular recombination model and $n_{eq} p_{eq}$ is the equilibrium value of the np product that assures that G_{npl} vanishes under equilibrium conditions. For DD theory with ideal gases all of this is simple because of the relationship $n_{eq} p_{eq} = n_i^2$ where n_i is the known intrinsic density. With a more general equation of state, this reduction is no longer possible, and it becomes necessary to solve the Poisson equation and therefrom to obtain n_{eq} and p_{eq} for inclusion in (2.2.17). The situation in DG theory is analogous, however, in this case one needs to solve three coupled differential equations (see below) for the equilibrium quantities.

2.3 Chemical potential formulation

In principle, the equations of Sects. 2.1 and 2.2 complete the formulation of DG theory. However, as with DD theory, under most circumstances these equations can be greatly simplified by a transformation to chemical potentials. This formulation is also important as the mathematical justification for band diagrams and for their extension to DG theory. For simplicity we focus on the electron gas whose stress tensor obeys (2.2.4) with (2.2.6) and (2.2.8), i.e.,

$$\tau^n = \left[-n^2 \frac{\partial \varepsilon_n}{\partial n} + 2n \nabla \cdot \left(n \frac{\partial \varepsilon_n}{\partial \Pi_n} \nabla n \right) \right] \mathbf{I} - 2n \frac{\partial \varepsilon_n}{\partial \Pi_n} \nabla n \nabla n \quad (2.3.1)$$

Motivated by the form of (2.1.2), we observe that

$$\begin{aligned}-\nabla \cdot \tau^n &= \nabla \cdot \left[n^2 \frac{\partial \varepsilon_n}{\partial n} - 2n \nabla \cdot \left(n \frac{\partial \varepsilon_n}{\partial \Pi_n} \nabla n \right) \right] \\ &\quad + 2 \nabla \cdot \left[n \frac{\partial \varepsilon_n}{\partial \Pi_n} \nabla n \nabla n \right]\end{aligned}$$

$$\begin{aligned}
&= n \nabla \left[\frac{\partial n \varepsilon_n}{\partial n} - 2 \nabla \cdot \left(n \frac{\partial \varepsilon_n}{\partial \Pi_n} \nabla n \right) \right] \\
&= q n \nabla \phi_n^{DG}
\end{aligned} \quad (2.3.2)$$

where the temperature and material properties have been assumed uniform, and ϕ_n^{DG} is a **generalized chemical potential** for the electron gas in DG theory that is defined by:

$$q \phi_n^{DG} \equiv \frac{\partial n \varepsilon_n}{\partial n} - \nabla \cdot \left(\frac{\Xi_n \nabla n}{n} \right) \quad (2.3.3)$$

The analogous development for holes clearly yields

$$\begin{aligned}
&-\nabla \cdot \tau^p = q p \nabla \phi_p^{DG} \\
&\text{where } q \phi_p^{DG} \equiv \frac{\partial p \varepsilon_p}{\partial p} - \nabla \cdot \left(\frac{\Xi_p \nabla p}{p} \right)
\end{aligned} \quad (2.3.4)$$

When the electron/hole momenta are negligible, the existence of the chemical potentials allows the linear momentum balance equations (2.1.2) to be replaced by much simpler integral forms, namely:

$$\begin{aligned}
\int_S \mathbf{n}(\phi_n^{DG} - \psi) dS &= \int_V \mathbf{E}_n dV \quad \text{and} \\
\int_S \mathbf{n}(\phi_p^{DG} + \psi) dS &= \int_V \mathbf{E}_p dV
\end{aligned} \quad (2.3.5)$$

In addition, in terms of the chemical potentials, the material response functions for linear gradient gases are readily obtained from (2.3.3) and (2.3.4) as

$$\begin{aligned}
\phi_n^{DG} &= \phi_n^{DD} - \frac{2}{s} \nabla \cdot (b_n \nabla s) \quad \text{and} \\
\phi_p^{DG} &= \phi_p^{DD} - \frac{2}{r} \nabla \cdot (b_p \nabla r)
\end{aligned} \quad (2.3.6)$$

where $\phi_n^{DD} \equiv \partial n \varepsilon_n^{DD} / \partial n$, $\phi_p^{DD} \equiv \partial p \varepsilon_p^{DD} / \partial p$, $s \equiv \sqrt{n}$, and $r \equiv \sqrt{p}$. The second terms in these expressions are sometimes called “quantum potentials” because of their formal similarity to a quantity in Bohm’s formulation of quantum mechanics [14].

2.4 Differential equations and boundary conditions

From the integral forms for the physical principles as given in Sects. 2.1 and 2.3, both differential equations and boundary conditions can be derived with their common source assuring consistency. The former are reached when the field variables are differentiable, usually via a direct application of the divergence theorem. Boundary conditions result when the field variables are not differentiable, and as in electromagnetism are usually derived by taking limits of the integrals over “Gaussian pillboxes”.

Charge/mass balance. The differential equations that follow from (2.1.1) are:

$$\frac{\partial n}{\partial t} + \nabla \cdot (n \mathbf{v}_n) = G_{npl} \quad \text{and} \quad \frac{\partial p}{\partial t} + \nabla \cdot (p \mathbf{v}_p) = G_{npl} \quad (2.4.1a)$$

and assuming no interface recombination, the corresponding boundary conditions are:

$$\begin{aligned}
\mathbf{n} \cdot [n^+ \mathbf{v}_n^+ - n^- \mathbf{v}_n^-] &= 0 \quad \text{and} \\
\mathbf{n} \cdot [p^+ \mathbf{v}_p^+ - p^- \mathbf{v}_p^-] &= 0
\end{aligned} \quad (2.4.1b)$$

Momentum balance. The differential equations that follow from (2.3.3) and (2.3.4) are:

$$\begin{aligned}
m_n \frac{d^n \mathbf{v}_n}{dt} &= q \nabla \Phi_n^{DG} + q \mathbf{E}_n - m_n \mathbf{v}_n \frac{G_{npl}}{n} \quad \text{and} \\
m_p \frac{d^p \mathbf{v}_p}{dt} &= -q \nabla \Phi_p^{DG} + q \mathbf{E}_p - m_p \mathbf{v}_p \frac{G_{npl}}{p}
\end{aligned} \quad (2.4.2a)$$

where $\Phi_n^{DG} = \psi - \phi_n^{DG}$ and $\Phi_p^{DG} = \psi + \phi_p^{DG}$ are generalized electrochemical potentials (or generalized quasi-Fermi levels) for DG theory, and the material (total) derivatives take their usual Eulerian forms of $d^n \mathbf{v}_n / dt \equiv \partial \mathbf{v}_n / \partial t + \mathbf{v}_n \cdot \nabla \mathbf{v}_n$ and $d^p \mathbf{v}_p / dt \equiv \partial \mathbf{v}_p / \partial t + \mathbf{v}_p \cdot \nabla \mathbf{v}_p$ as in fluid mechanics. In words these equations say that the gradients of the electrochemical potentials act as driving “forces” on the carrier gases and are balanced by drag and/or by inertial “forces”. The boundary conditions expressing momentum balance are:

$$\begin{aligned}
\phi_n^{DG+} - \phi_n^{DG-} &= f_n \quad \text{and} \\
\phi_p^{DG+} - \phi_p^{DG-} &= -f_p
\end{aligned} \quad (2.4.2b)$$

where f_n and f_p are the forces per charge exerted by the semiconductor surface or interface on the carriers and we have used the fact that the electric potential is continuous (assuming no surface dipoles).

Electrostatics. The differential equations of electrostatics as derived from (2.1.3) are familiar:

$$\nabla \cdot \mathbf{D} = q(N - n + p) \quad \text{and} \quad \mathbf{E} = -\nabla \psi \quad (2.4.3a)$$

and the usual boundary conditions are:

$$\begin{aligned}
\mathbf{n} \cdot [\mathbf{D}^+ - \mathbf{D}^-] &= \sigma \quad \text{and} \\
\mathbf{t} \cdot [\mathbf{E}^+ - \mathbf{E}^-] &= 0 \quad \text{or} \\
[\psi^+ - \psi^-] &= 0
\end{aligned} \quad (2.4.3b)$$

where σ is the surface charge density and \mathbf{t} is the vector tangent to the interface.

Energy balance. These equations are not given explicitly here, however, they were used in deriving (2.2.1).

2.5 Microscopic connections

The near-universal view among electronics researchers and engineers is that the fundamental justification for DD theory—and by implication, for DG theory—is *necessarily* microscopic and must be grounded in the Boltzmann equation, the Wigner-Boltzmann equation or other such formulas [15]. This view is belied by the material presented in Sects. 2.1–2.4 that in itself constitutes a purely *macroscopic* foundation. Known as a classical field theoretic approach [1], the power of this strategy was first revealed early in the 19th century by Euler and Cauchy who used it to obtain a correct macroscopic theory of solids (elasticity) while knowing very little about the underlying microscopics. That this macroscopic approach has been enormously consequential in many areas of mathematical physics is indisputable.⁶ At the same time, its existence in no way implies that microscopic approaches are without value, and indeed the two perspectives should be regarded as complementary with each having advantages and each shedding light on the basic physics.

The main advantage of a microscopic development of a macroscopic description lies in the possibility of establishing explicit connections between macroscopic coefficients in the material response functions (e.g., the mobility) and the underlying microscopics (e.g., the scattering physics). Such connections are invaluable for enhancing physical understanding, and they can sometimes be quantitative and especially useful for projecting the performance of semiconductors that have not yet been grown of device quality [16]. Given the limited length of this review, our discussion of the microscopic viewpoint will be confined to the crucial DG equation of state (2.3.6).

Although not originally developed in the context of a transport theory, derivations of the DG equation of state go back to the beginnings of quantum mechanics (as noted in Sect. 1.2) and to work of Weizacker [7]. A variety of other derivations with differing assumptions followed [17]; below we sketch a fairly general development based on density functional theory that is due to Perrot [17]. The starting point is Mermin's proof [18] that there exists a functional of the density $n(\mathbf{x})$, namely $G[n]$, that is independent of the potential $V(\mathbf{x})$ and for which

$$\Omega[n(\mathbf{x})] = \int V(\mathbf{x})n(\mathbf{x})d\mathbf{x} + \frac{e^2}{2} \int \frac{n(\mathbf{x})n(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} d\mathbf{x}d\mathbf{x}' + G[n]$$

with $G[n] = \int g[n]d\mathbf{x}$ (2.5.1)

⁶Other prominent examples where correct macroscopic equations were obtained by macroscopic methods before the relevant microscopic physics was understood are fluid dynamics (inviscid case by Euler beginning in 1752; viscous case by Navier in 1837), metallic conduction (Drude in 1900), superconductivity (London in 1935) and liquid crystals (Ericksen in 1964).

is a minimum when $n(\mathbf{x})$ is the equilibrium density. At this minimum, Ω is the grand potential of the system, the first integral in (2.5.1) is the potential energy associated with $V(\mathbf{x})$, the second integral is the energy associated with the Coulomb interaction, and the local functional $g[n]$ groups contributions from the kinetic energy and from exchange and correlation.

The derivation begins by considering situations in which the density is slowly varying (but with possibly large excursions) so that $g[n]$ is well represented by a gradient expansion [9]:

$$g[n] = g_0(n) + g_2^{(2)}(n)\nabla n \cdot \nabla n + g_4^{(2)}(n)\nabla^2 n \nabla^2 n + g_4^{(3)}(n)\nabla^2 n \nabla n \cdot \nabla n + g_4^{(4)}(n)(\nabla n \cdot \nabla n)^2 + \dots \quad (2.5.2)$$

whose form has been restricted by rotational invariance and by the idea that it can be unique only to within a divergence. Inserting (2.5.2) into (2.5.1) and minimizing the grand potential $\Omega[n]$ subject to the constraint that the average density is n_0 (handled with a Lagrange multiplier μ) leads to the condition

$$V(\mathbf{x}) - \mu + e^2 \int \frac{n(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} d\mathbf{x}' + g_0'(n) - g_2^{(2)'} \nabla n \cdot \nabla n - 2g_2^{(2)} \nabla^2 n + \dots = 0 \quad (2.5.3)$$

By identifying $\mu = -q\Phi_n^{DG}$, $g_0(n) = n\varepsilon_n(n)$, and $g_2^{(2)}(n) = eb_n/2n$, it is readily shown that (2.5.3) is the same as the equilibrium version of (2.4.2a)₁ with (2.3.6)₁. Hence the foregoing constitutes a microscopic derivation of the DG equation of state (2.3.6)₁. More importantly, we can obtain explicit microscopic formulas for the coefficients in (2.5.2)—and especially for $g_2^{(2)}$ since it relates to the density-gradient coefficient b_n —by examining (2.5.3) for the case of an almost constant (but possibly rapidly varying) density. Considering a uniform electron gas with an imposed small positive charge perturbation of wave number \mathbf{q} , it is readily shown from (2.5.3) that the electronic polarizability can be written to lowest order as [9]

$$\alpha(\mathbf{q}) \equiv 1 - \frac{g_0''}{4\pi} q^2 + \left[\left(\frac{g_0''}{4\pi} \right)^2 - \frac{g_2^{(2)'}}{2\pi} \right] q^4 + \dots \quad (2.5.4)$$

where $q \equiv |\mathbf{q}|$. Since the polarizability is related to the electric susceptibility $\chi(\mathbf{q})$ by $4\pi\chi(\mathbf{q})/q^2 = \alpha(\mathbf{q})/[\alpha(\mathbf{q}) - 1]$ the coefficients in (2.5.4) can be estimated in the random phase approximation via Lindhard's famous formula:

$$\chi(\mathbf{q}) = -\frac{e^2 m}{4\pi^3} \int \frac{f_{\mathbf{k}-\mathbf{q}/2} - f_{\mathbf{k}+\mathbf{q}/2}}{\hbar^2 \mathbf{k} \cdot \mathbf{q}} d\mathbf{k}$$

where $f_{\mathbf{k}} \equiv \frac{1}{1 + e_{\mathbf{k}}}$ and $e_{\mathbf{k}} \equiv \exp\left[\frac{1}{k_B T} \left(\frac{\hbar^2 k^2}{2m} - \mu \right)\right]$ (2.5.5)

Focusing on slowly varying perturbations (to which the Lindhard expression best applies), we expand (2.5.5) for small \mathbf{q} and obtain

$$\chi(\mathbf{q}) = -\frac{e^2}{4\pi^3} \int \frac{\partial f_{\mathbf{k}}}{\partial \mu} \left[1 + \frac{\hbar^2 q^2}{8mk_B T} (1 - 2f_{\mathbf{k}} e_{\mathbf{k}}) + \frac{1}{6} \left(\frac{\hbar^2 \mathbf{k} \cdot \mathbf{q}}{2mk_B T} \right)^2 (1 - 6f_{\mathbf{k}} e_{\mathbf{k}} + 6f_{\mathbf{k}}^2 e_{\mathbf{k}}^2) \right] d\mathbf{k} \quad (2.5.6)$$

where the right side can be reduced to Fermi-Dirac integrals $F_n(\mu/k_B T)$. By comparing terms in (2.5.4) and (2.5.6), we can then find expressions for g'_0 and $g_2^{(2)}$, and therefrom microscopic formulas for φ_n and b_n :

$$\varphi_n = g'_0 = k_B T \ln \left(\frac{n}{N_C} \right) \quad \text{and} \quad (2.5.7a)$$

$$b_n = 2ng_2^{(2)} \equiv \frac{\hbar^2}{4em_n r_n} \quad (2.5.7b)$$

where

$$N_C = 2 \left(\frac{m_n k_B T}{2\pi^2 \hbar^2} \right)^{3/2} \quad \text{and} \quad r_n = \frac{3F_{-1/2}^2}{F_{1/2} F_{-3/2}} \quad (2.5.7b)$$

It is readily shown [19, 20] that if the electrons could be in a pure state (i.e., at absolute zero with the Pauli exclusion principle ignored), then the factor r_n would be unity.⁷ Hence, r_n can be said to represent the effect of the statistics. According to the above derivation, r_n depends only on $\mu/k_B T$ and in the non-degenerate limit ($\mu \rightarrow -\infty$), (2.5.7b)₂ shows directly that $r_n = 3$, whereas in the degenerate limit ($\mu \rightarrow \infty$), Weizacker's result is recovered ($r_n = 9$) [7]. Between these limits, numerical calculations reveal a smooth transition as plotted in Fig. 2.5.1.

By providing the formulas in (2.5.7a) and (2.5.7b), the foregoing derivation does indeed represent a microscopic foundation for DG theory that provides insight into the origin and meaning of the DG coefficient. This is of genuine value. At the same time, it should be emphasized that the approximations/assumptions in the derivation are severe, and it is unknown how the results change when, for instance, the density variations are both large and rapid as they are in the situations of most interest. More generally, this points up a generic shortcoming of microscopic derivations, that they usually provide *necessary* conditions for a particular result, and not the *sufficient* conditions that one would really like to have.

Lastly, we note that when DG theory is used as a phenomenology (see Sects. 3.4 and 4.5) the coefficient b_n is used as

⁷The formula (2.5.7b)₂ does not apply to this case because the expansions on which it is based become invalid at low temperature as is evident from (2.5.6).

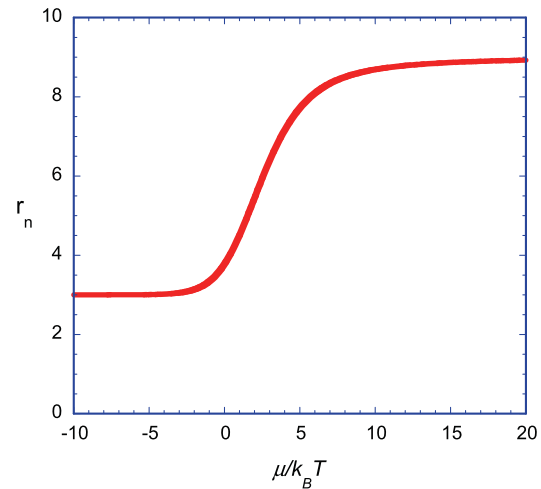


Fig. 2.5.1 Plot of the DG statistical factor r_n as a function of the normalized chemical potential as calculated microscopically in the random phase approximation

a regression parameter with fits being made either to quantum mechanical calculations or to experiments. When comparing with quantum mechanics, one generally knows the effective mass, and so the fitting can be regarded as a means of determining the statistical factor r_n . If the fits are instead made to experiment, it seems more convenient to assume $r_n = 3$, and then use the fitting to estimate a DG effective mass m_n^{DG} .

3 Quantum confinement

3.1 DG-confinement theory: physics, mathematics, and numerics

For many important semiconductor devices the quantum physics of evanescence (see Sect. 1.2) is manifested as an equilibrium or quasi-equilibrium phenomenon. Most commonly this occurs in “quantum wells” wherein the (quasi-) equilibrium is imposed by confining potential barriers in one-, two- or three-dimensions. For example, in 1D quantum wells form the channels of most field-effect transistors and the active layers of heterostructure lasers; in 2D they are the FINFETs and nanowires currently of considerable research interest; and in 3D they are the semiconductor quantum dots that are attractive as luminescent biolabels, and possibly also for optoelectronic devices and solar cells. The equilibrium nature of such situations in the confined direction(s) implies that components of the inertia in those directions can be neglected. If additionally one can assume that the transport in any non-confined direction(s) is scattering-dominated, then as in DD theory, the momentum terms in (2.4.2a) can be

neglected and these equations reduce to force (per charge) balance equations:

$$-\nabla\Phi_n^{DG} = \mathbf{E}_n \quad \text{and} \quad \nabla\Phi_p^{DG} = \mathbf{E}_p \quad (3.1.1)$$

where $\Phi_n^{DG} = \psi - \phi_n^{DG}$ and $\Phi_p^{DG} = \psi + \phi_p^{DG}$ are again the generalized electrochemical potentials of DG theory. Being relevant to transport in confined situations, we refer to (3.1.1) as the defining equations of **DGC** (for DG-Confinement) **theory**. They are direct generalizations of the analogous equations of DD theory in which the gradients of electrochemical potentials act as driving “forces” on the carrier gases and are balanced by drag forces [2, 3, 13]. Introducing the assumptions of an ideal gradient gas (2.2.13b) and linear drag (2.2.16) we obtain

$$\mathbf{J}_n = n\mu_n\nabla\psi - D_n\nabla n + 2\mu_nb_nn\nabla\left(\frac{\nabla^2 s}{s}\right) \quad \text{and} \quad (3.1.2)$$

$$\mathbf{J}_p = -p\mu_p\nabla\psi - D_p\nabla p + 2\mu_pb_pp\nabla\left(\frac{\nabla^2 r}{r}\right)$$

where the diffusion coefficients are given by the Einstein relations, e.g., $D_n \equiv \mu_n \partial \phi_n^{DD} / \partial n$. The equations in (3.1.2) are identical in form to the DD current equations except for the addition of DG quantum correction terms. That these correction terms arise from the electron gas response means that they are of the same physical character as the normal diffusion terms, and are thus properly viewed as quantum diffusion currents. Nevertheless, the mathematics is such that as in DD theory one can also interpret the gradients of the electrochemical potentials as “effective electric fields”, in which case the DG effects can be construed as Bohmian “quantum potentials” [14].

For mathematical/numerical purposes it is helpful to recast the DGC differential equations (2.4.1a), (3.1.1) with (2.3.6), and (2.4.3a) as the 2nd-order system

$$\nabla \cdot (n\mu_n \nabla \Phi_n^{DG}) = G_{npl} - \frac{\partial n}{\partial t} \quad \text{and} \quad (3.1.3a)$$

$$\nabla \cdot (p\mu_p \nabla \Phi_p^{DG}) = \frac{\partial p}{\partial t} - G_{npl}$$

$$\nabla \cdot (b_n \nabla s) = \frac{s}{2} [\Phi_n^{DG} - \psi + \phi_n^{DD}] \quad \text{and} \quad (3.1.3b)$$

$$\nabla \cdot (b_p \nabla r) = \frac{r}{2} [\Phi_p^{DG} - \psi - \phi_p^{DD}]$$

$$\nabla \cdot (\varepsilon_d \nabla \psi) = q(n - p - N) \quad (3.1.3c)$$

to be solved for Φ_n^{DG} , Φ_p^{DG} , s , r and ψ . Much like the comparable equations of DD theory (i.e., those obtained from (3.1.3a), (3.1.3b) and (3.1.3c) by setting b_n and b_p to zero), these equations form an elliptic-parabolic PDE system and this has a variety of mathematical and numerical conse-

quences.⁸ For example, a simple scaling of the equations reveals that they are characterized by five intrinsic length scales: the Debye screening length, electron and hole diffusion lengths, and electron and hole quantum lengths with the latter being of the form $L_Q = \sqrt{b/\phi}$ (where b is either b_n or b_p , and ϕ is a voltage/energy scale such as a barrier height). That all of these length scales multiply high-order derivatives of (3.1.3a), (3.1.3b) and (3.1.3c) means that they represent singular perturbations [23, 24]. Because of the elliptic/parabolic character of the equations, the characteristic “breakdown” in the solutions that occurs when these intrinsic lengths are “small” (i.e., compared to the geometry) will be localized into boundary layers. The boundary layers associated with the diffusion and Debye lengths are familiar, while that of L_Q defines the layer in which the quantum influence of the confining barrier is manifested. Because the electron and hole gases are incoherent, the L_Q are of the order of the deBroglie wavelength of an individual electron, and as this is usually on a scale of nanometers, the quantum boundary layers will generally be nested well inside the other layers.

In solving the DGC equations numerically a variety of approaches can be considered.⁹ This author has found effective a “Slotboom” approach [25] wherein the densities are transformed according to $u = k_B T \ln(n/N_0)/2q$ and $v = -k_B T \ln(p/N_0)/2q$ with N_0 being an arbitrary density. In terms of the numerical variable set Φ_n^{DG} , Φ_p^{DG} , u , v , and ψ , the equations are:

$$\nabla \cdot (e^{qu/k_B T} \mu_n \nabla \Phi_n^{DG}) = \frac{G_{npl}}{N} - \frac{q e^{qu/k_B T}}{k_B T} \frac{\partial u}{\partial t} \quad \text{and} \quad (3.1.4a)$$

$$- \nabla \cdot (e^{-qv/k_B T} \mu_p \nabla \Phi_p^{DG}) = \frac{q e^{-qv/k_B T}}{k_B T} \frac{\partial v}{\partial t} + \frac{G_{npl}}{N}$$

$$\nabla \cdot \left(\frac{2qb_n}{k_B T} e^{qu/k_B T} \nabla u \right) = e^{qu/k_B T} [\Phi_n^{DG} - \psi + \phi_n^{DD}] \quad \text{and} \quad (3.1.4b)$$

$$\nabla \cdot \left(\frac{2qb_p}{k_B T} e^{-qv/k_B T} \nabla v \right) = -e^{-qv/k_B T} [\Phi_p^{DG} - \psi - \phi_p^{DD}]$$

$$\nabla \cdot (\varepsilon_d \nabla \psi) = qN \left(\exp\left(\frac{2qu}{k_B T}\right) - \exp\left(-\frac{2qv}{k_B T}\right) - 1 \right) \quad (3.1.4c)$$

where we have assumed uniform doping and set $N_0 = N$. A practical algorithm results when these equations are dis-

⁸ Some mathematically oriented discussions of the DGC equations have appeared in [21–23].

⁹ For special cases, [19] developed some semi-analytical approaches based on a variational principle and a first energy integral that exist for the system (3.1.3a), (3.1.3b) and (3.1.3c) in equilibrium. There have also been efforts to use the approximation techniques of singular perturbation theory [24]. These possibilities are of academic interest, but for practical work the purely numerical approach is almost always preferred because of its flexibility and scope.

cretized in either a finite-difference or finite-element framework, with Newton's method invoked to handle the nonlinearities and with the linear algebra solved by a direct approach. For added efficiency, a Scharfetter-Gummel discretization scheme may be implemented [26–28]. As in the analogous DD situation, an essential for getting this scheme to converge is a good initial guess; the author has found that starting with artificially inflated values for the Debye and quantum lengths makes for a robust strategy.

Tools for device analysis using the DGC equations as described above are readily available. All of the major commercial device simulators offer a DGC capability, e.g., it comes as an option in Silvaco's ATLAS simulator,¹⁰ in Synopsys's SENTAURUS simulator,¹¹ and in Xilinx's ISE Simulator [29].¹² The author has also found Comsol's general-purpose finite element simulator¹³ to be quite effective at solving these equations, although at present this tool is incapable of implementing the Scharfetter-Gummel discretization.

3.2 Quantum wells in 1D

One-dimensional quantum wells are not only critical elements of many semiconductor devices, but they also provide a parade ground for exhibiting some of the basic characteristics of the DG approach, and a testbed for gauging the range and accuracy of DGC theory. Facilitating matters is the fact that the corresponding quantum mechanical analyses—typically in the effective-mass Schrödinger or Hartree (Schrödinger-Poisson) approximations—are usually straightforward both in formulation and in computation. Indeed, quantum mechanics often constitutes a practical alternative for such problems. In this regard it should be remembered that the main basis for interest in DG theory is not simple 1D quantum wells, but rather its broader utility for practical device engineering applications with larger multi-dimensional structures, non-equilibrium conditions, and/or transport that is strongly coupled to electrostatic (or other) fields.

The first careful study of the DGC equations as applied to various 1D quantum well problems appeared in [19]. A second relevant paper focused on quantum confinement in silicon inversion layers with direct relevance to MOS technology [30]. The latter paper also explored several generalizations of the linear version of DGC theory of Sect. 3.1 aimed at extending its range. Selected findings from these two investigations are presented in this section, along with results from similar unpublished calculations that have been performed more recently.

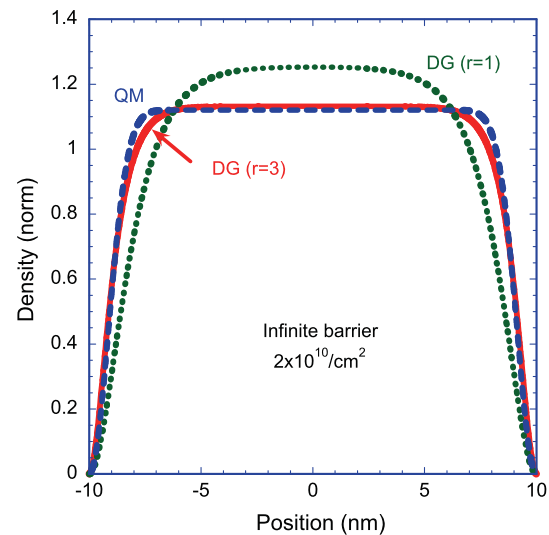


Fig. 3.2.1 Electron density profiles in an infinite barrier quantum well as computed by DGC theory with $r_n = 1$ or $r_n = 3$ and compared with quantum mechanical calculations. For this 20 nm wide well with $m_n = 1.0m_e$, $T = 300$ K, and an areal density of $2 \times 10^{10} \text{ cm}^{-2}$, the assumption of $r_n = 3$ works quite well

We begin by considering a simplest case of quantum wells in which electrons are confined by barriers that are infinitely high. In this idealized arrangement, the infinite barrier forces the density to zero at the edges of the well, and so only the region inside the well needs to be modeled. As a first simulation we test the validity of DGC theory by comparing it with quantum mechanics for the case of a 20 nm quantum well at room temperature with $m_n = 1.0m_e$ and a well density of $2 \times 10^{10} \text{ cm}^{-2}$. For the quantum mechanical calculation one of course solves a one-dimensional particle-in-box problem for the eigenvalues E_i and normalized eigenfunctions $\psi_i(x)$, and then computes the electron density in equilibrium with band-filling according to the Fermi-Dirac distribution. For the DGC calculations, we assume the statistical factor r_n introduced in Sect. 2.5 takes the theoretical values of either the pure-state ($r_n = 1$) or the high-temperature limit ($r_n = 3$). From the solution profiles plotted in Fig. 3.2.1 it is evident that the latter limit is the relevant one, with the agreement being quite good although not perfect. And in general, similar comparisons show DGC theory with $r_n = 3$ to provide an excellent representation of the confinement so long as the effective mass is not too small, the quantum well is not too narrow, and/or the temperature is not too low. As illustration, in Fig. 3.2.2 we show density profiles across half of the same symmetric 20 nm well for various values of the effective mass. DGC theory with $r_n = 3$ again does well until the mass becomes quite small ($< 0.1m_n$). A similar set of curves with varying temperature in Fig. 3.2.3 finds that appreciable error enters only at the very lowest temperatures (< 20 K). The discrepancies in these plots can be understood as occurring

¹⁰See <http://www.silvaco.com>.

¹¹See <http://www.synopsys.com>.

¹²See <http://www.xilinx.com>.

¹³See <http://www.comsol.com>.

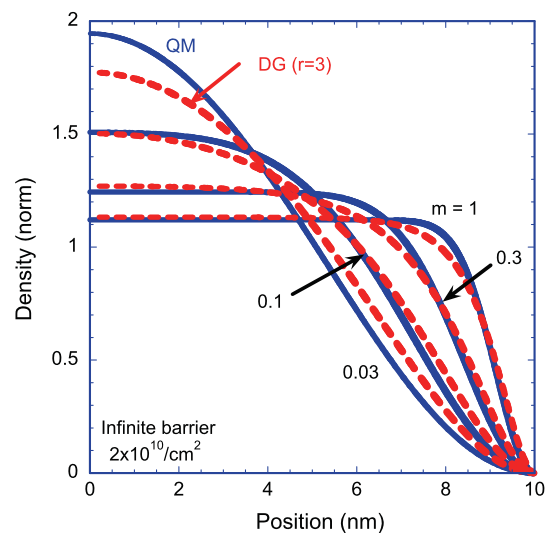


Fig. 3.2.2 Density profiles across half the quantum well width with all as in Fig. 3.2.1 except that the effective mass is varied

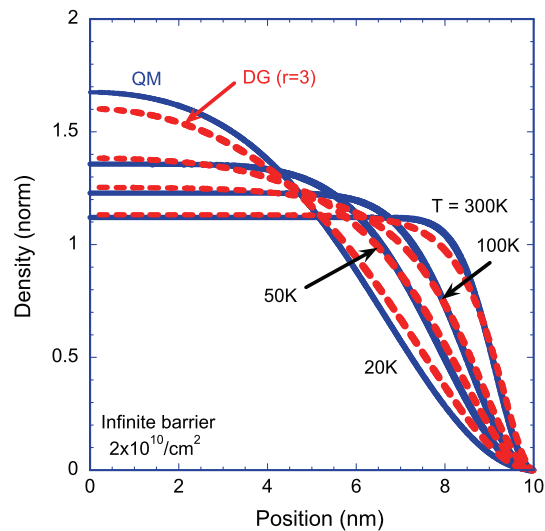


Fig. 3.2.3 A similar plot to Fig. 3.2.2 but with temperature as the parameter being varied

when the mass, temperature, and/or well thickness are such that only a few subbands are occupied and “high temperature” average statistics no longer represents a good approximation [19]. In the smallest mass and lowest temperature curves in Figs. 3.2.2 and 3.2.3 where the errors are largest, roughly 80% of the electrons are in the lowest subband (see Fig. 3.2.4). Finally, examining the extreme limit of a 2 nm quantum well in which only a single sub-band has appreciable occupation, we find in Fig. 3.2.5 that DGC theory in the pure-state limit ($r_n = 1$) provides a somewhat more accurate representation. Thus, at moderate densities the most challenging cases for DGC theory are those in which the electrons are concentrated in a few sub-bands. One approach for such problems would be to construct a multi-subband the-

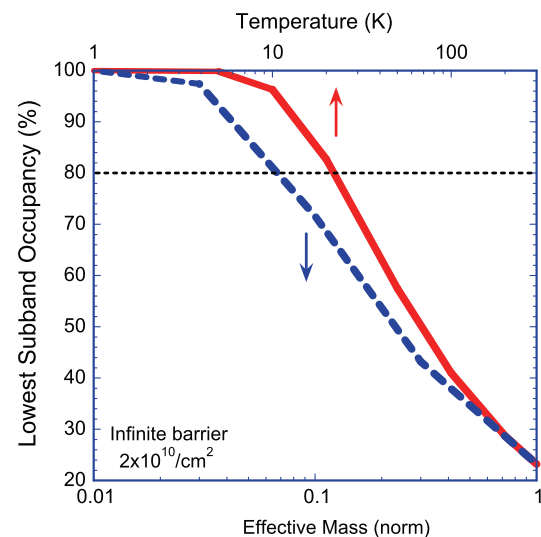


Fig. 3.2.4 The occupancy of the lowest subband in the quantum well situations of Figs. 3.2.1–3.2.3 as the effective mass or the temperature are varied

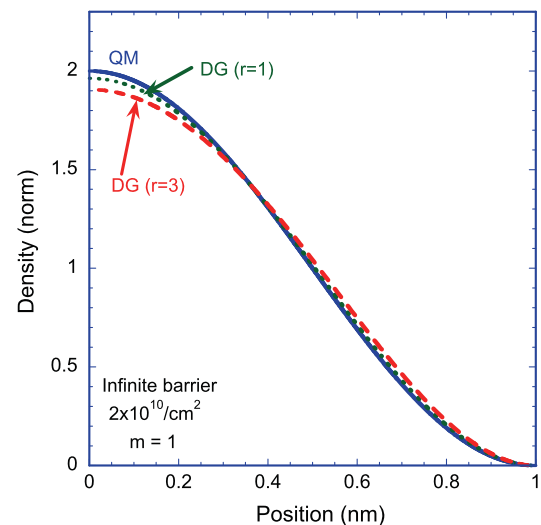


Fig. 3.2.5 A similar plot to Fig. 3.2.1 but for a 2 nm quantum well. In this case, DGC theory with $r_n = 1$ is seen to perform slightly better

ory in which each subband satisfies its own DGC equations [11, 30]. However, given the cumbersomeness of such a theory and the fact that it would have to be calibrated with a quantum mechanical simulation, a better approach is probably that of a “single-gas” DGC phenomenology as discussed in Sect. 3.3.

The inadequacy with respect to subband structure is not the only kind of error that can occur in DGC theory’s description of quantum confinement. A second important error is in the representation of the inhomogeneous electron gas as the density becomes elevated and higher-order gradient effects become more important. A mild manifestation of this type of error was already seen in the high curvature regions

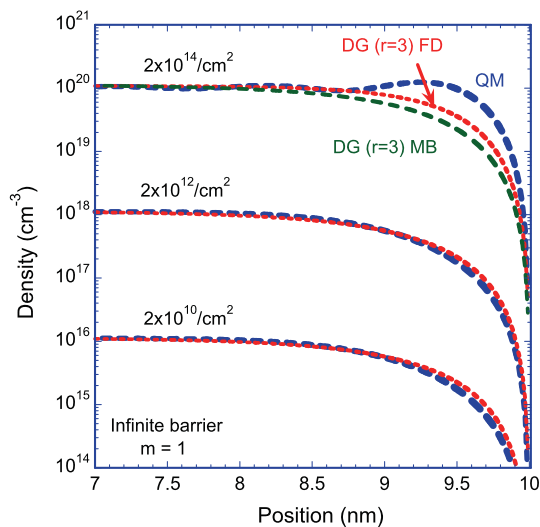


Fig. 3.2.6 Density profile comparisons between DGC theory ($r_n = 3$) and quantum mechanics at high density. DGC curves are shown both with and without quantum compressibility effects included

of the DGC profile in Fig. 3.2.1 ($r_n = 3$ curve). To study the issue under more extreme circumstances, we simulate the same quantum well but with higher levels of sheet density as shown in Fig. 3.2.6. For such modeling it should be noted that one needs to include the effect of quantum compressibility on $\phi_n^{DD}(n)$ in (3.1.1), although Fig. 3.2.6 shows the size of this correction to be small. In any event, the main point of the comparisons in Fig. 3.2.6 is that when the sheet density gets as high as $2 \times 10^{14} \text{ cm}^{-2}$ (which corresponds to volumetric densities of $\sim 10^{20} \text{ cm}^{-3}$), not only do the quantitative discrepancies get larger, but the DGC description completely fails to represent the Friedel oscillations that are a well-known aspect of quantum mechanical screening [31]. Whether a higher-order DGC theory could capture this latter phenomenon is not known, but in any case it is clear that linear DGC theory is significantly deficient when it comes to representing confined electron gases at high density.

From a technological perspective, the most important quantum well is that that occurs in silicon MOSFETs adjacent to oxide interfaces. Because the barrier presented by SiO_2 to the conduction band is quite high ($\sim 3.5 \text{ eV}$), electron inversion layers on p-type silicon can be regarded as existing in quantum wells with effectively infinite barriers. However, this infinite-barrier situation is more complicated than those discussed previously because, for the usual (100) orientation material, the six degenerate conduction bands of silicon split into two non-equivalent sets of valleys with different masses and capacitance characteristics. Nevertheless, in [30] it was found that DGC solutions with $r_n = 3$ and m_n taking the bulk Si value ($0.328m_e$) agree remarkably well with those obtained from Hartree simulations. As illustration, in Fig. 3.2.7 we show electron density profiles as computed by both DGC theory and quantum mechanics

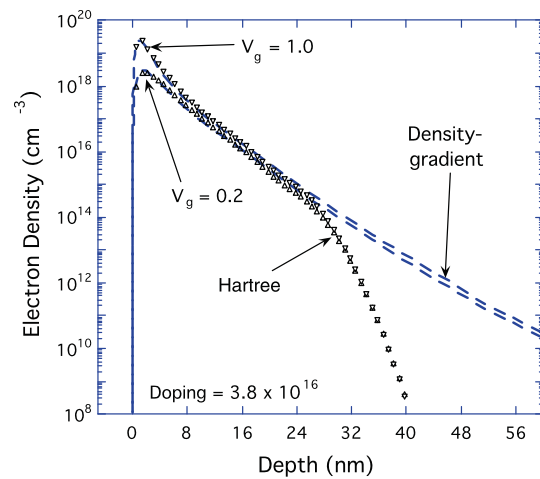


Fig. 3.2.7 Comparison of predictions of DGC theory and quantum mechanics for a silicon inversion layer. The disagreement far from the surface is due to the quantum mechanical calculation not including enough subbands

for two different gate voltages in inversion, and find excellent agreement in the important high-density region near the Si– SiO_2 interface. Curiously, this agreement does not hold up far from the surface where the density is very low, an error that is actually in the quantum mechanics and is due to the 80 subbands included in that calculation not being enough. Thus the much simpler DGC theory seems to be providing the more accurate description. But a closer examination [30] reveals that the good performance of DGC theory is fortuitous and arises from compensating errors that keep the product $m_n r_n$ in (2.5.7a)₂ roughly constant and about equal to $0.328 \times 3 \cong 1$. Specifically, as the device is biased into stronger inversion, increased occupancy in the lighter-mass in-plane minima causes the average effective mass m_n to drop, while at the same time the increased band filling causes the factor r_n to rise. A better macroscopic approach to this physics, introduced in [30] but not discussed further here, was based on noting that the changing distribution of carriers between the non-equivalent valleys can be modeled as a nonlinear DG effect.

Because of the degree of control afforded by theory, comparing theoretical predictions is especially effective for learning about the physical content of the theories. But semiconductor device physics is still largely an experimental discipline, and therefore the ultimate test of theory remains how well it incorporates the physics germane to a particular device, while leaving out extraneous details so as to enhance understanding and utility.¹⁴ To emphasize this point, we briefly review results from an investigation that compared DG theory with experiments on ultra-thin oxide MOS

¹⁴This statement applies not only to DGC theory but also to quantum mechanics. In general, the latter is better at bringing in all the physics, but less adept at focusing on the essentials.

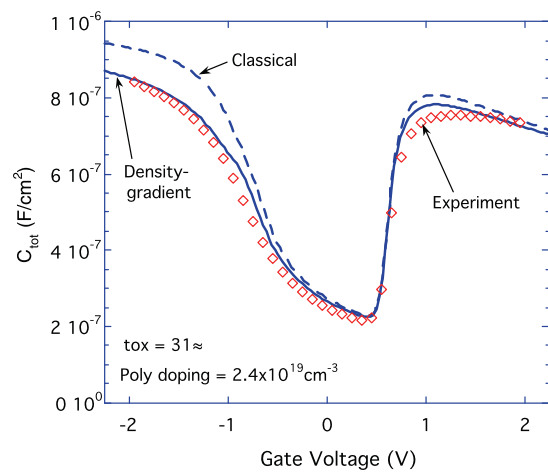


Fig. 3.2.8 Comparison between the experimental C–V characteristics of an ultra-thin-oxide silicon capacitor and predictions made by DD and DGC theories. The agreement of the latter is reasonable, though not perfect

capacitors [32]. One complication of these devices was that they had polysilicon gates and therefore could exhibit depletion effects not just in the semiconductor substrate but also in the gate.¹⁵ In Fig. 3.2.8 we compare the experimental capacitance-voltage characteristics from one such capacitor with estimates obtained by both DD and DGC theories. The classical DD calculation is clearly incorrect in that it shows none of the capacitance reduction associated with quantum effects. DGC theory obviously does much better, although it is not perfect. That there is good agreement in accumulation (i.e., negative bias) where the capacitance is most influenced by the quantum effects shows that the discrepancies in the DGC curve come not from flaws in the quantum representation but rather from other inadequacies such as in describing generation/recombination.¹⁶

To this point we have discussed confinement problems in which the barriers are well approximated as being infinitely high. But of course this is an idealization and in many practical semiconductor device situations the finiteness of the barrier height plays an important role. Having a finite barrier brings in two additional physical phenomena, namely, the thermal excitation of carriers over the barrier (“thermionic emission”) and the exponentially decaying penetration of carriers into the barrier via quantum evanescence. Both of these effects allow charge carriers to enter the barrier region and both can be important in devices, e.g., allowing trapping/detrapping in the barrier material. In a

¹⁵Another complication was that the substrate had an inhomogeneous doping profile that had to be carefully profiled using SIMS and included in the simulations.

¹⁶Threshold voltages are hard to predict with any theory, and so the matching of theoretical and experimental threshold voltages in Fig. 3.2.8 was achieved by curve-fitting.

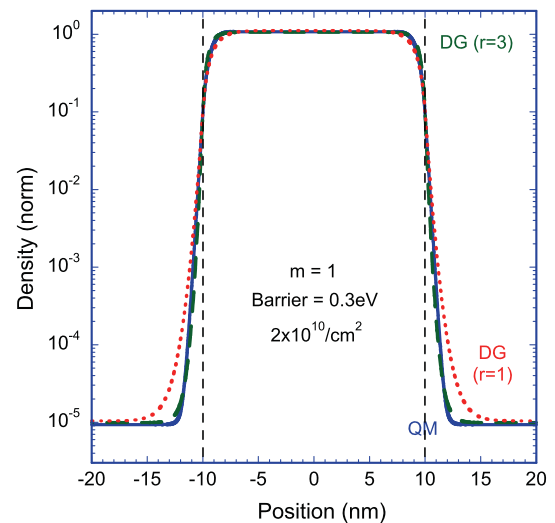


Fig. 3.2.9 A similar plot to Fig. 3.2.1 but with a finite barrier height of 0.3 eV

quantum mechanical description both phenomena are captured (for equilibrium) so long as the continuous spectrum of eigenstates at high energy is also included. Macroscopically, DD theory was deficient in providing only a good description of the thermionic emission, but DGC theory would seem to make a simple unified treatment possible. In particular, the thermionic emission is included in the foregoing equations (just as in DD theory) by having the ordinary chemical potential(s) be discontinuous by the band offset(s), and the physics of evanescence is represented by the DG term. That all of this occurs within a single description is elegant, parsimonious, and potentially useful. However, it is also wrong! What is erroneous is that within the barrier the (higher energy) electrons participating in the thermionic emission form a largely separate population from those electrons (of lower energy) that evanesce into the barrier, yet the unified DGC description treats them as one. As a practical matter, this can often be justified by noting that in most situations one or of the other of these two phenomena dominates. Alternatively, one can properly capture the underlying physics by splitting the electron gas inside the barrier in two, an idea pursued further below.

To illustrate the DG simulation of carrier confinement by barriers that are finite in height we consider a situation analogous to that of Fig. 3.2.1 with a 20 nm quantum well and $m_n = 1$, but now with a barrier height of 0.3 eV. DGC results with $r_n = 1$ and $r_n = 3$ are shown in Fig. 3.2.9 together with the quantum mechanical profile, and we observe that again DGC theory with $r_n = 3$ provides the better representation. But looking at these profiles in greater detail in Fig. 3.2.10, we find that neither version of DGC theory (labeled “w/diff”) is fully capturing the barrier penetration behavior with the quantum mechanical decay being essentially a simple exponential (i.e., a straight line in the semi-log plot)

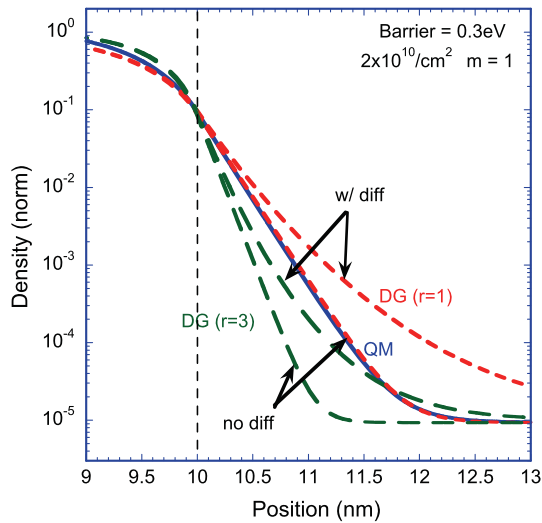


Fig. 3.2.10 A semi-log version of Fig. 3.2.9 emphasizing the barrier penetration and comparing DGC solutions with and without diffusion included in the description of the evanescent carriers inside the barrier

whereas both DGC results are sub-exponential. The curvature in the DGC profiles arises from the fact that the unified description unphysically permits the evanescent carriers to diffuse. With this in mind, we introduce the “split” description mentioned in the previous paragraph. In this treatment, thermionic emission is incorporated via a second barrier population whose density is determined simply by the chemical potential at the barrier edge, while the evanescent population is treated as before but with ordinary diffusion turned off and ϕ_n^{DD} fixed at its value at the barrier edge. The results of this simulation, again for both $r_n = 1$ and $r_n = 3$ and with the densities of the two barrier populations added together, are also shown in Fig. 3.2.11. As expected, the new DGC profiles are indeed now simple exponentials. Moreover, we now find that the DGC calculation with $r_n = 1$ is the preferred one being almost identical inside the barrier to that predicted by quantum mechanics. The reason for this limit being appropriate is that, as a function of depth, the barrier penetration is increasingly dominated by the longest wavelength (lowest energy) carriers, and this implies a suppression of the carrier statistics effect that is responsible for larger r_n values (see Sect. 2.5). The remaining error in the $r_n = 1$ profile is largely the result of the previously discussed inaccuracy of the $r_n = 1$ theory inside the well. This is highlighted in a linear plot (not shown) that again finds DGC theory with $r_n = 3$ to provide a much better representation inside the well. To get the best of both the obvious solution is to create a “hybrid” that uses $r_n = 1$ in the barrier and $r_n = 3$ in the well. And this works very nicely if one smoothes the transition between the r_n values.

Based on our earlier discussion, it is no surprise that the results for the finite barrier quantum well degrade when the effective mass decreases, the well narrows, the temperature drops or the density becomes very high. As an illus-

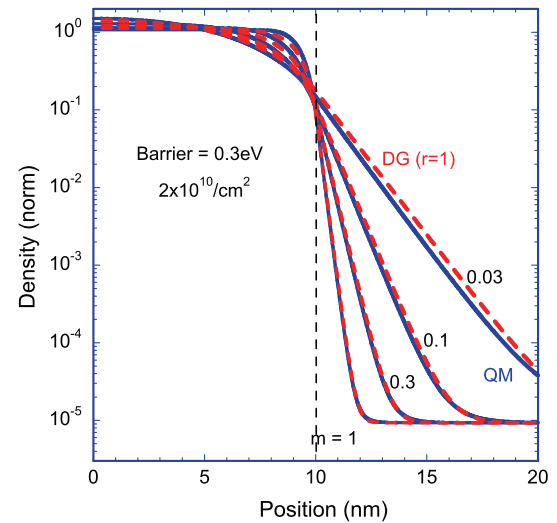


Fig. 3.2.11 The analogous plot to Fig. 3.2.2 for a quantum well with a finite barrier. The DGC solution assumes $r_n = 1$ and leaves out the unphysical diffusion of the evanescent carriers inside the barrier

tration of this, in Fig. 3.2.11 we test DGC $r_n = 1$ predictions against quantum mechanics for masses of $m_n = 1.0$, 0.3, 0.1, and 0.03. The agreement in this semi-log plot is remarkably good in all cases with the larger error in the barrier for the lightest mass mostly being due to propagation of the increased error inside the well. Similar calculations with narrower wells, lower temperatures or higher densities also show growing errors just as were seen in Figs. 3.2.3 and 3.2.6.

3.3 DGC phenomenology

Establishing the precise boundary between physics and curve-fitting is generally challenging, and in our case this means it is hard to know just when the DGC description ceases to be a legitimate field theory and instead becomes a phenomenology. On the one hand, as noted in Sect. 2.2, it can be tricky distinguishing between an inadequacy in the particular choices for the material response functions [e.g., in (2.2.13b)] and a true failure of the DG framework itself. And on the other hand, because of their manifest limitations, microscopic calculations like those of Sect. 2.5 tend not be a reliable guide, e.g., the lack of agreement in the previous section between DGC theory with $r_n = 1$ or $r_n = 3$ is at best circumstantial evidence for the inadequacy of DGC theory. With these remarks as background, we observe that an important characteristic of DGC theory (and also of DG tunneling theory as discussed in Sect. 4) is that it often allows for remarkably accurate curve-fits to quantum mechanical density profiles [19]. Why this has never been explained, but its broad-ranged accuracy suggests that it may be more

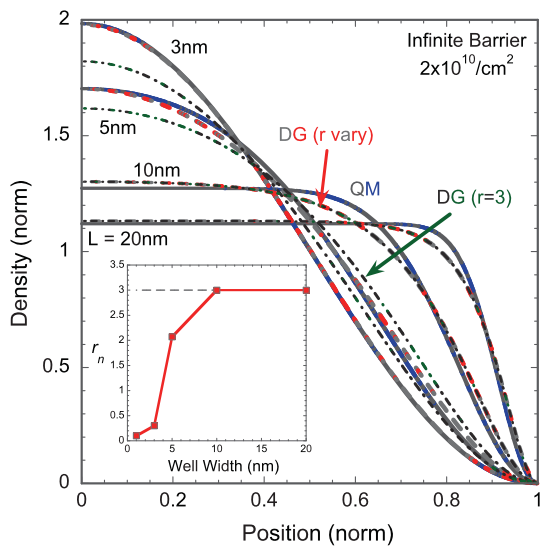


Fig. 3.3.1 Phenomenological use of DGC theory to fit quantum mechanical results for infinite barrier quantum wells of varying width. The inset shows the values of r_n need to produce these fits

than just “lucky” mathematics.¹⁷ In any event, because it involves curve-fitting of microscopic results (and for other reasons outlined below), we refer to this approach as *DGC phenomenology*. The basic procedure is no more than to use the DG effective mass ($m_n^{DG} \equiv m_n r_n$) or one of its constituent parameters (m_n or r_n) as a coefficient for fitting the quantum mechanics. And rather than employ a formal regression procedure, we do the fits “by eye” while imposing certain “reasonable” constraints as explained below.

As an example of DGC phenomenology, we examine in Fig. 3.3.1 simulation results for infinite barrier quantum wells of various widths as obtained by quantum mechanics, by DGC $r_n = 3$ theory, and by fitting solutions of the DGC equations using r_n as the parameter. For the latter, as the well becomes wide, one finds the “best” value of r_n to be close to 3, with the exact choice affecting merely the *distribution* of the error. Because there seems little physical meaning in such fine-tuning, we impose the constraint that r_n be no more than three. The results shown in Fig. 3.3.1 were obtained with the r_n value varying with well width as shown in the inset. The fits are clearly not perfect, but overall the performance is quite good. Of most interest is the fact that, for narrow wells, using a reduced value of r_n greatly improves the entire profile. As noted in the previous paragraph, this variation in r_n is not a proof that we are now beyond the capabilities of *any* DG-based field theory. But the observed direct dependence of r_n on the size of the “box” the electron gas resides in shows that this is a non-local effect and thus is an improper extension of *linear* DGC field theory as

we have formulated it. Thus the fits in Fig. 3.3.1 must be regarded as phenomenology.

Further discussions of the phenomenological uses of DG theory appear in Sects. 3.4 and 3.5 and in Sect. 4 in the contexts of various more complicated device situations that involve not just ideal equilibrium confinement but also transport.

3.4 Quantum confinement in multi-dimensions

The analyses of quantum-confined situations in multi-dimensions using either quantum mechanics or DGC theory are directly analogous to those in one-dimension, although of course more numerically demanding. For this reason, we limit coverage of this topic to just two examples of cylindrically symmetric quantum dots (QDs) embedded in barrier material. In both cases we assume the electrostatics can be neglected, and that the QDs are charged with a single electron and are at a reduced temperature (77 K) in order to limit the number of eigenvalues (40) needed for quantum mechanics to represent the Fermi-Dirac occupancy accurately. The first illustration is of a simple cylinder, with a radius of 10 nm, a height of 20 nm, a barrier height of 0.7 eV, and unit electron mass. For the quantum mechanical calculation only bound states have been included while in the DGC calculation the effect of the continuum is automatically incorporated (thermionic emission), but this difference is insignificant since the barrier height is of appreciable size. Also, in the DGC calculation the charging of the QD (by the assumed single electron) is set by the choice of the constant electrochemical potential. In Fig. 3.4.1 we compare the density profiles as computed by quantum mechanics and by DGC theory, with the latter assuming that, as in Fig. 3.2.10, that in the QD $r_n = 3$ (solid line) or $r_n = 1$ (dotted line) and in the barrier $r_n = 1$ and ϕ_n^{DD} is constant for the evanescent component of the total carrier density. Figure 3.4.1a gives the profiles in the radial direction from the QD center, while Fig. 3.4.1b has the profiles along the polar axis again from the QD center. The data of Fig. 3.4.1a is also shown in semilog form in Figs. 3.4.2a and 3.4.2b. Clearly, as in the analogous 1-D situations, DGC theory with $r_n = 3$ is in much better agreement with quantum mechanics than is the $r_n = 1$ theory. The $r_n = 3$ theory disagrees primarily in again not capturing the Friedel oscillations that in the radial direction are especially pronounced because of the importance of quantum mechanical states with significant angular momentum in the cylindrical geometry. When the shape of the dot is changed, the importance of these states can be reduced or magnified leading to smaller or larger perturbations, and with DGC theory giving a correspondingly better or worse representation. As in 1-D, when the masses, the QD size, or the temperature are reduced, DGC theory will become increasingly less accurate. However, just as discussed in

¹⁷Some discussion of a possible physical basis appeared in the last reference in [16].

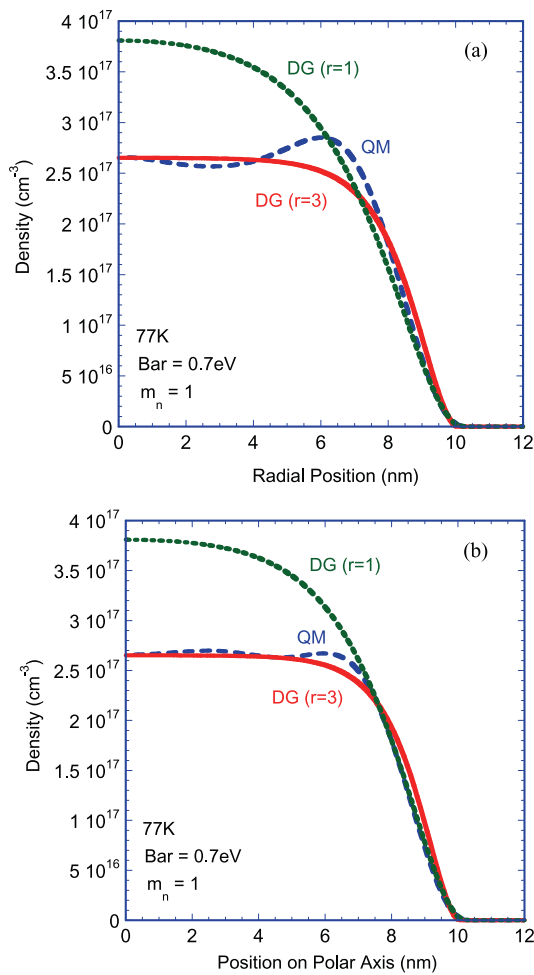


Fig. 3.4.1 Comparison of density profiles as computed by quantum mechanics and DGC theory with $r_n = 3$ or $r_n = 1$ for a cylindrical QD. The profiles are along cutlines (a) in the radial direction and (b) along the polar axis both with origin at the QD center

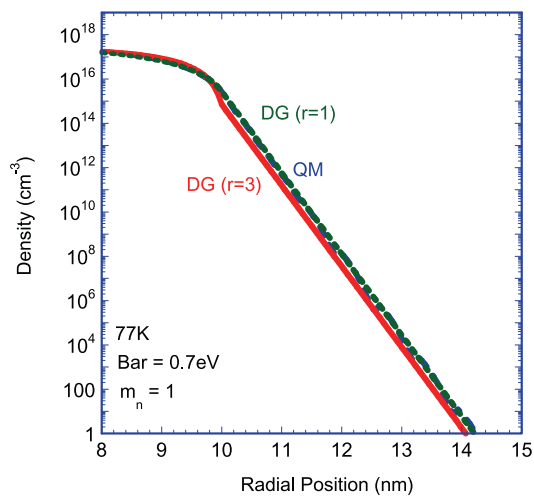


Fig. 3.4.2 Semilog version of the plot in Fig. 3.4.1a; the analogous plot to Fig. 3.4.1b looks much the same

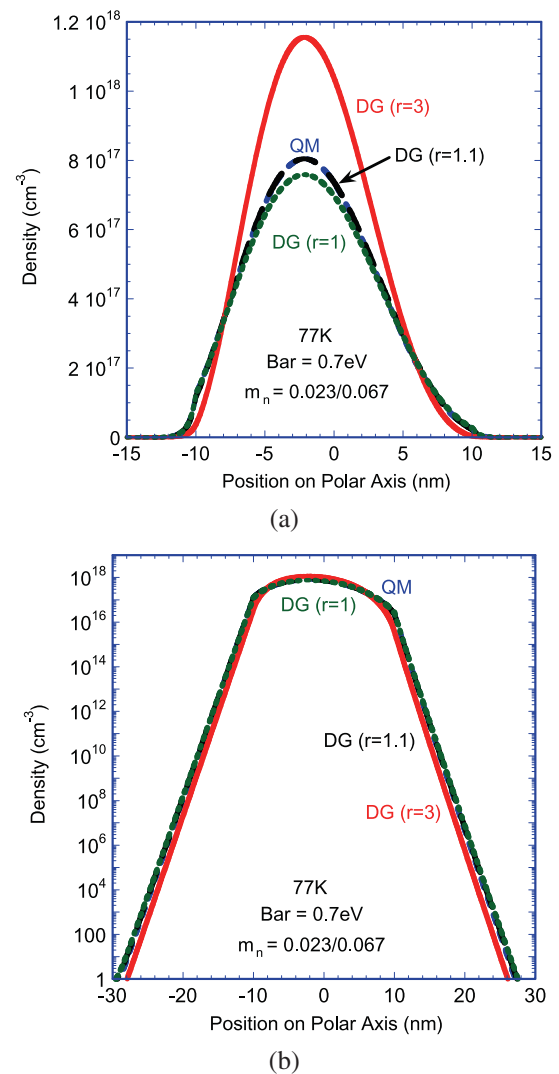


Fig. 3.4.3 Comparison of density profiles as computed by quantum mechanics, by DGC theory with $r_n = 3$ and $r_n = 1$, and by a DGC phenomenology with $r_n = 1.1$ for a half-ellipsoid QD composed of InAs and with the barrier material being GaAs. The (a) linear and (b) semilog profiles are plotted along the polar axis with the asymmetry about the origin arising from the asymmetry of the QD

Sect. 3.3, it turns out that the multi-dimensional theory can often serve as the basis for a remarkably accurate phenomenology simply by using r_n or m_n^{DG} as a fitting parameter. To illustrate, we examine the case of an asymmetrical (in the polar direction) half-ellipsoid with a major axis of 20 nm in the polar direction and a minor axis of 10 nm in the radial direction and with the materials having masses corresponding to an InAs QD ($0.023m_e$) and a GaAs ($0.067m_e$) barrier. In Figs. 3.4.3a and 3.4.3b we show linear and semilog profiles along the polar axis as computed by quantum mechanics, by DGC theory with $r_n = 3$ or $r_n = 1$, and by a DGC phenomenology in which r_n is taken to have a best-fit value of 1.1 inside the QD. The asymmetry in all of the solutions is due to the QD being a half-ellipsoid, and the agreement in the case of the DGC phenomenology is seen to be superb.

Overall, it should be said that the quality of the multi-dimensional DGC solutions for the QDs is very good. Interestingly, in our examples the results seem unaffected by the fact that the dots are occupied by just a single electron. This is again due to the “smearing” effect of quantum mechanics that renders the continuum assumption of DG theory quite robust. On the other hand, it must be acknowledged that the DGC calculations do not provide much of the detailed information that is of great interest in QD applications such as the level spacings. Thus, the DGC treatment of these cases seems of limited practicality and should be regarded mostly as illustration of the range and fidelity of the theory/phenomenology for multi-dimensional problems in which quantum-confinement is important.

3.5 Confined transport in quantum wells

As discussed in Sect. 3.1, when one has quasi-equilibrium quantum confinement in the transverse direction(s) and scattering-dominated transport in the non-confined lateral direction(s), the inertia terms will be *negligible in all directions* and the DGC equations in (3.1.1) will constitute the relevant formulation. This “orthogonal” description of the physics is often apropos for modeling FETs, whether the confinement is in 1D as in an ordinary planar structure, or in 2D as in nanowire FETs.¹⁸ The earliest such application of a DGC-like theory to FET simulation was in [34–36]. We begin this section by reviewing a much more recent transistor example where again the physics can reasonably be treated with the “orthogonal” approach.

Recent research on the antimony-containing semiconductors of InSb, GaSb, and their alloys has boosted the hole mobility measured in FET channels to record values of as high as 1500 cm²/V-s, and this has encouraged the dream of a high performance CMOS technology based on III–V semiconductors. As part of an effort to investigate this possibility, the author recently employed DGC theory to look at the performance potential of III–V pFETs [37]. One such device, studied in depth experimentally in [38], had a 5 nm InSb quantum well channel flanked by Al_{0.35}In_{0.65}Sb barriers. A similar GaSb design studied in [39] had a 7.5 nm GaSb quantum well with AlAs_{0.25}Sb_{0.75} barriers. The hole density profiles across these two structures as computed by quantum mechanics in the $k \cdot p$ approximation for various levels of charging of the wells are shown in Figs. 3.5.1a and 3.5.1b. Because of the narrowness of these wells and the splitting of the heavy- and light-hole valence bands (analogous to the conduction band splitting seen in the Si inversion layer in Sect. 3.2 but due both to the confinement

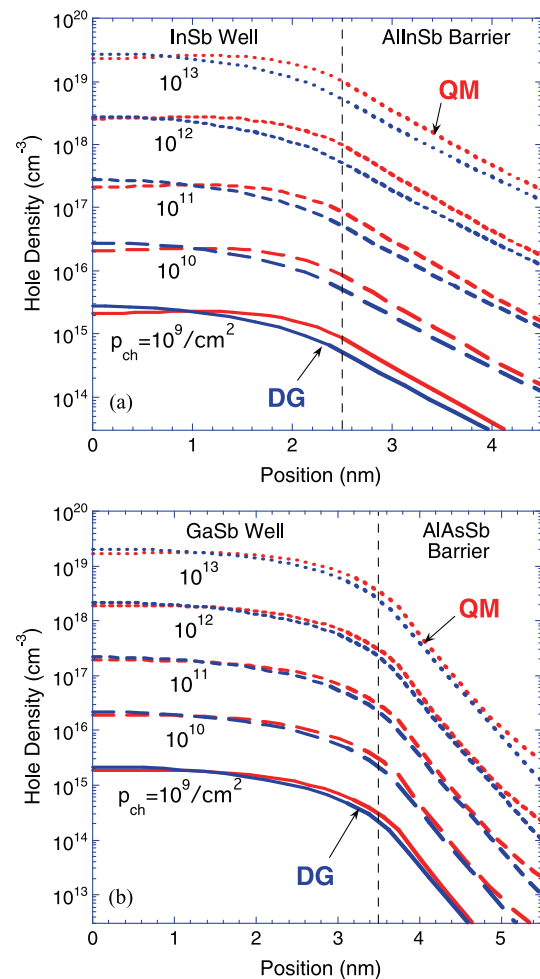


Fig. 3.5.1 Comparison of density profiles as computed by quantum mechanics and has fit by DGC theory over a wide range of sheet densities for (a) 5 nm InSb quantum well with Al_{0.35}In_{0.65}Sb barriers and (b) 7.5 nm GaSb quantum well with AlAs_{0.25}Sb_{0.75} barriers. The good quality fits were obtained with m_p^{DG} values of $0.04m_e$ and $0.06m_e$, respectively

and mechanical strain), we employ the phenomenological DGC approach as discussed in Sect. 3.3. We find that the quantum mechanical profiles in Figs. 3.5.1a and 3.5.1b can each be well fit with single values of the fitting parameter m_p^{DG} — $0.04m_e$ and $0.06m_e$ for the InSb and GaSb pFETs, respectively—used for all levels of charging. As in earlier comparisons, the results are not perfect but are remarkably good considering the tightness of the geometry, the complexity of the valence bands, and the wide density variations; the better agreement in the GaSb case is presumably due to its wider well, heavier mass, and larger band offset. Next, we assume a simple mobility model with the measured low-field mobilities, velocity saturation and surface scattering [37], and then compute the heterostructure FET characteristics for the geometry shown in the inset to Fig. 3.5.2. The two results plotted in Fig. 3.5.2 are for InSb and GaSb devices with 40 nm channel lengths, and without and with recessed

¹⁸Many investigators have been guided by a similar logic in creating hybrid “theories” (sometimes referred to as a quantum drift-diffusion approach) that marry a 1D Hartree analysis across the channel with DD transport descriptions along the channel [33].

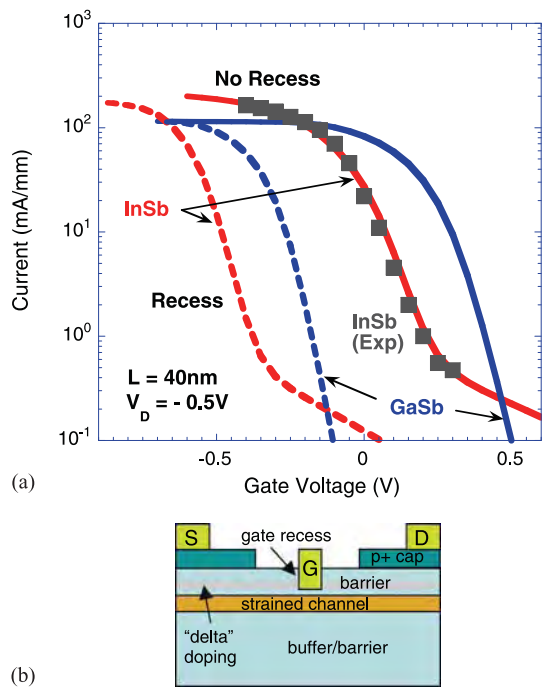


Fig. 3.5.2 DGC-computed transfer characteristics for 40 nm InSb and GaSb FETs both with and without recessed gates. Also shown are experimental result from [38] for a 40 nm InSb device with a non-recessed gate. The inset depicts the device structure with the gate recess

gates. The former mimics the InSb experiment of [38], and as seen in the figure, excellent agreement is obtained. The leakage current in this device is due to source-drain leakage [38], and the good fit of Fig. 3.5.2 is obtained by assuming a mobility in the AlInSb barrier material of $50\text{ cm}^2/\text{V}\cdot\text{s}$. Going further, we then used these device models to project the benefits of device scaling. For a recessed-gate design, in Fig. 3.5.3 we find that as the devices are scaled, the InSb devices are about three times faster than the GaSb devices, but the latter devices have two orders of magnitude less standby power (assuming gate leakage is negligible). A similar examination of InGaSb channel devices shows them to come close to combining the best of both the binary channel devices [37]. One final point regarding these simulations is to verify that the component of the DG term in (3.1.2)₂ along the channel is small so that it is indeed a good assumption to treat the physics as “orthogonal”. To this end, in Fig. 3.5.4 we plot the I–V curves of a non-recessed 20 nm InSb FET in which the component of the mass tensor along the channel is varied. Clearly, when the mass is small enough ($< 0.01m_e$) there is an effect, but because the mass is surely not this low, we conclude that the physics along the channel direction is indeed well represented by ordinary DD theory and the “orthogonal” DGC treatment is thus validated.

If similar calculations to those of Fig. 3.5.4 were carried out for devices with channel lengths much shorter than 20 nm the component of the DG term in (3.1.2)₂ along

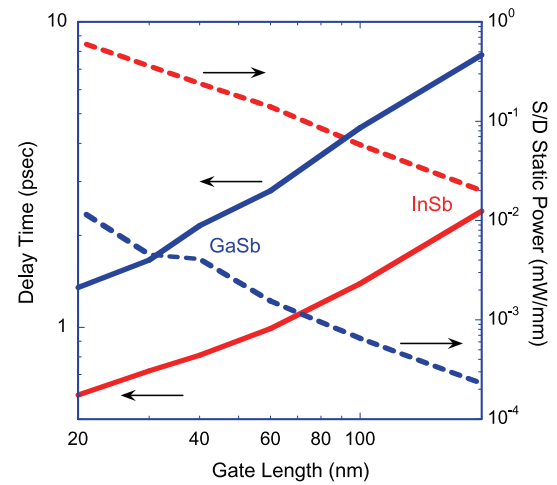


Fig. 3.5.3 Scaling projections of the gate delay and static power dissipation of InSb and GaSb FETs as estimated by DGC simulation

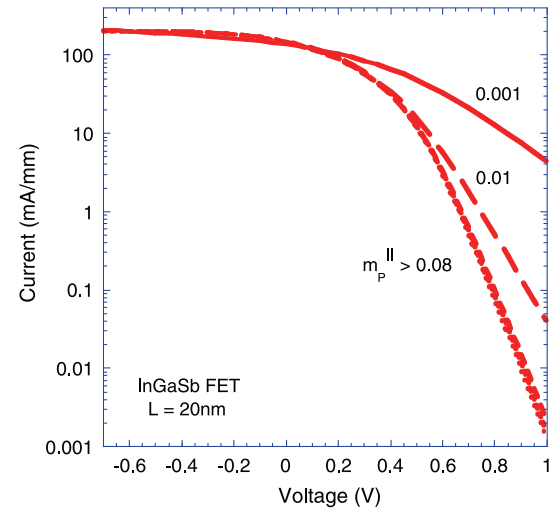


Fig. 3.5.4 DGC-calculated transfer characteristics for a 20 nm InSb pFET with the lateral effective mass for holes treated as a parameter

the channel direction would undoubtedly become significant and the “orthogonal” basis for DGC theory would be undermined. It is nevertheless worth asking whether there is any meaning and/or value in such a DGC description where the lateral component of the DG term produces significant current. That such currents can arise in a FET model¹⁹ was first noted in [40], and in that paper it was further suggested that they could serve as a way of modeling source-drain tunneling under subthreshold conditions. From the standpoint of the physics this idea suffers from the same flaw seen in our earlier treatment of barrier penetration in Sect. 3.2 (see Fig. 3.2.10), namely that these “tunneling” carriers would exhibit an unphysical normal diffusion. Alternatively, an error is signaled by the fact that the computed source-drain

¹⁹That DGC theory admits scattering-dominated “tunneling” currents was first reported in [40] in the context of 1D problems.

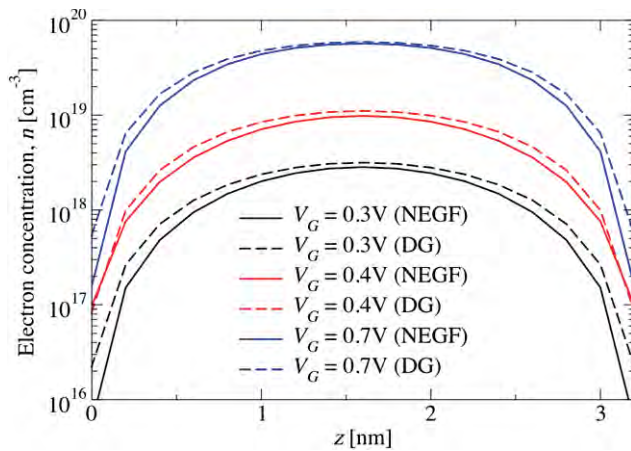


Fig. 3.5.5 Density profiles from [42] across the width of a nanowire FET as computed by NEGF and the DGC phenomenology with varying transverse mass. The best DGC fit results when $m_{nT}^{DG} \cong 0.067$

“tunneling” would be proportional to the material’s mobility. But despite this error, DGC theory might still be of value as an easy way of including source-drain tunneling in a DD-like simulation as pointed out in [40]. This is akin to representing the physics of Fig. 3.2.10 by the unphysical but reasonably accurate DGC (“w/diff”) $r_n = 3$ curve in that figure. The potential value of this simple approach for engineering is best appreciated by comparing it with the quantum mechanical alternative that becomes nearly intractable if various real-world complications such larger-scale multi-dimensional geometry, fully coupled electrostatics, and/or scattering are included.²⁰ Although not nearly as complex, the proper DG approach introduced in Sect. 3.2 (and discussed in much greater detail in Sect. 4) would also present challenges. The cost of the simple DGC phenomenology is that, as demonstrated in [40], quantitative accuracy requires that it be calibrated with quantum mechanics and/or experiments.

As a second and more challenging application of the DGC phenomenology to source-drain tunneling, the Glasgow group recently discussed similar modeling of a Si nanowire transistor [42]. About as small as a “normal” FET can possibly be, their device had a gate length of only 4 nm, a wire cross-section of 3.2×3.2 nm, and a S/D doping of 10^{20} cm^{-3} . This ultra-scaled FET is surely well beyond the scope of DG theory and a proper description must necessarily be microscopic, with the non-equilibrium Green’s function (NEGF) method being used in [42]. To explore how well a DGC phenomenology might handle this extreme problem, [42] took the DG effective mass to be a second-rank tensor [see (2.2.13b) with (2.5.7a)] with differing components in the transverse and longitudinal directions for ex-

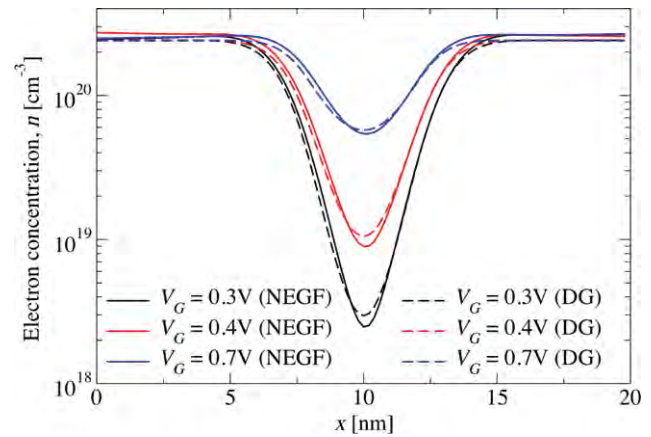


Fig. 3.5.6 Density profiles from [42] along the length of a nanowire FET as computed by NEGF and the DGC phenomenology with varying lateral mass. The best DGC fit results when $m_{nL}^{DG} \cong 0.1$

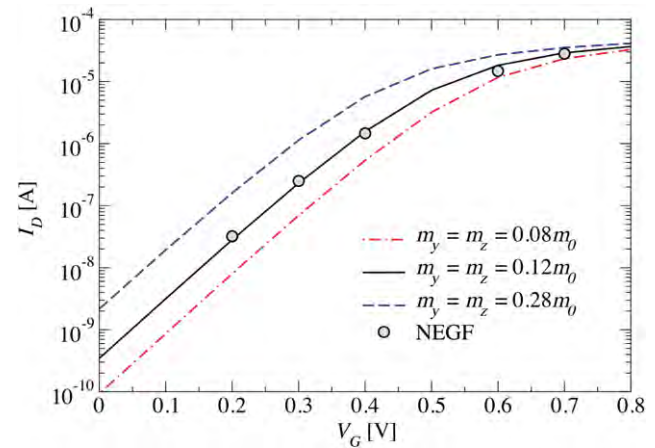


Fig. 3.5.7 Subthreshold I–V characteristics from [42] for a nanowire FET as computed by NEGF and the DGC phenomenology with varying lateral mass. The best DGC fit results when $m_{nL}^{DG} \cong 0.27$

tra flexibility. The transverse mass was used to fit the quantum confinement in the nanowire much as in Sect. 3.2 and with similarly impressive results. Figure 3.5.5 shows density profiles across the nanowire width with the optimal transverse mass²¹ of $m_{nT}^{DG} = 0.067m_e$ giving very good agreement. As seen in the figure, they found the quality of these fits to be relatively independent of both the gate voltage and the longitudinal mass. The situation in the longitudinal direction is less clean because the physics one is trying to represent is more complicated, having both source-drain tunneling and electrostatic confinement of the electrons within the source and drain regions. And as seen in Figs. 3.5.6 and 3.5.7, the results of [42] do indeed seem to reflect this complexity: They find that different longitudinal masses are needed to fit the NEGF-calculated density profiles along the

²⁰Because of the complexity of this calculation, to the author’s knowledge it has been carried out only in very small structures in the “ballistic” limit using NEGF.

²¹Because of the definition in (2.5.7b), our masses are one third of those given in [42].

channel (Fig. 3.5.6, $m_{nL}^{DG} = 0.1m_e$) and the subthreshold leakage current (Fig. 3.5.7, $m_{nL}^{DG} = 0.27m_e$). In any event, given that this is a “molecular” scale FET, these results obtained using the simple DGC phenomenology still seem extraordinary.

4 Quantum tunneling

In most semiconductor transport situations, strong scattering causes the charged carrier gases to behave very differently from similar gases in the vacuum, e.g., in plasma physics or electron optics. For instance, in classical electron transport it is the dominance of scattering that allows inertia to be neglected in Newton’s Second Law, and which thereby leads to the DD transport equations. And in Sect. 3 we saw corresponding quantum transport examples that were governed by the scattering-dominated DGC equations. Nevertheless, there are many semiconductor transport situations where weak-scattering conditions prevail. These generally occur under circumstances of strong scaling when the geometric size becomes small compared to the mean free path, the so-called Knudsen regime of gas dynamics. Short-channel “ballistic” transistors are a prime example involving classical transport, and for these situations a ballistic transport description in which inertia plays a key role is essential. For quantum transport, weak-scattering conditions are again commonly encountered when the transport distance is small, and this is best exemplified by quantum tunneling in the elastic regime. The subject of this section is the “ballistic” version of DG theory that describes quantum transport when scattering is negligible (and inertia again plays a significant role as we shall see) and whose most important application is barrier tunneling.

In contrast to the scattering-dominated DGC description of Sect. 3, the DG approach to quantum tunneling as discussed in this section is little used by the device research and development communities at present. One reason for this, as emphasized in Sect. 1, is the apparent contradiction in using a “classical” theory to describe a “quantum” phenomenon. Again this is possible because the DG description is macroscopic, and so is classical only in the sense of conserving momentum and energy of the *population* and not of *individual electrons* (as a microscopic theory that is classical would do). Another concern that has been broached in the literature is computational stability [43]. The main goal of this section is to review the equations, interpretation, performance and limitations of DG tunneling theory in hopes of increasing its acceptance and use. Thus our goal is not to cover all possible tunneling situations, but rather to highlight the evidence that DG tunneling theory is legitimate as a theory, and potentially of great value as an engineering tool. The clearest understanding of DG tunneling to date appeared in a recent paper [44] and the treatment here is based primarily on that work.

4.1 DG-tunneling theory: physics, mathematics, and numerics

Given the extremely short path lengths involved (typically 1–3 nm), quantum tunneling usually takes place without appreciable scattering, a circumstance generally referred to as “elastic tunneling”.²² The appropriate form of DG theory for this regime is therefore the equations of Sect. 2 with scattering (and recombination) terms neglected. We call this DG formulation **DGT** (for DG-Tunneling) **theory**, and we observe that just as DGC theory is DD theory with quantum corrections, DGT theory can be viewed as ballistic transport theory with quantum corrections. Assuming the DG correction is that of a linear gradient gas (2.2.13a), the governing differential equations of DGT theory can then be written as:

$$\nabla \cdot (n\mathbf{v}_n) = -\frac{\partial n}{\partial t}, \quad \nabla \cdot (p\mathbf{v}_p) = \frac{\partial p}{\partial t} \quad (4.1.1a)$$

$$m_n \frac{d^n \mathbf{v}_n}{dt} = q \nabla \left[\psi - \phi_n^{DD} + \frac{2}{s} \nabla \cdot (b_n \nabla s) \right] \quad (4.1.1b)$$

$$m_p \frac{d^p \mathbf{v}_p}{dt} = -q \nabla \left[\psi + \phi_p^{DD} - \frac{2}{r} \nabla \cdot (b_p \nabla r) \right] \quad (4.1.1c)$$

where the left hand sides in (4.1.1b) involve the material derivatives that were defined in connection with (2.4.2a). A further implication of the narrowness of the barriers is that space charge will almost always be negligible (except when the electrodes are metal [45]) so that the right side of (4.1.1c) can usually be ignored. As in ballistic transport, the lack of scattering also leads to complications. For one thing, the negligible interaction among carriers implies that the electron/hole populations injected from different electrodes must each be described by their own transport equations. Furthermore, the directionality of the flows means that the physics—and hence the boundary conditions—at the emitting (or “upstream”) electrode will differ from that of the absorbing (or “downstream”) electrode.

Central to an understanding of DGT theory is an appreciation of the import of the left sides of (4.1.1b). These terms are of course Newtonian representations of the fact that the electron gas has (effective) mass/inertia/kinetic energy. At the same time, as discussed in Sect. 2.5, the DG terms in these equations bring in a lowest-order representation of the quantum mechanical kinetic energy operator. It is a fundamental assertion of DG theory that having both of these terms is *not double counting*, but rather is a statement that the total kinetic energy can be broken up into macroscopic and microscopic contributions much like the split in

²²The important topic of inelastic tunneling will not be considered in this review.

classical gas dynamics of microscopic kinetic energy into a macroscopic inertia term and a pressure term that represents the aggregate effect of particle motion about the mean. Whether or not such a break-up is valid for quantum transport is pivotal for establishing whether the DG description of tunneling is actually physically meaningful, and is not simply either curve-fitting or quantum mechanics in disguise.

Regarding the last point, there is indeed a superficial similarity between the DG equations and quantum mechanics, and especially with the pure-state Schrödinger equation when put in its well-known “hydrodynamic” form [20]. However, the application to tunneling is most revealing of the disparate natures of these two theories that sit on opposite sides of the microscopic-macroscopic dichotomy (see Sect. 1.1). One major difference already noted is the treatment of kinetic energy and its splitting in the macroscopic approach. Another crucial difference, discussed further in Sect. 4.2, is the essentiality of boundary conditions in the macroscopic approach. In quantum mechanics, boundary conditions are not fundamental—indeed the very idea of separating two materials by a sharp mathematical boundary/interface is a macroscopic one. A third difference between the theories is mathematical and has to do with the nature of the quantum mechanical “perturbation” set by the small parameters \hbar^2 or b_n (or b_p). In both cases, the parameter multiplies the highest-order derivative and hence is referred to as a singular perturbation. But as is well known in quantum mechanics, the Schrödinger equation breaks down globally as \hbar vanishes [46], whereas DG theory breaks down only locally in the limit of vanishing b_n (or b_p). The upshot is that DG theory’s quantum effects are restricted to localized boundary layers.

Equations (4.1.1a), (4.1.1b) and (4.1.1c) are written in a form appropriate for multi-dimensions, and for some problems this generality is clearly critical, e.g., tunneling in an STM [47]. However for most device tunneling problems, a 1D (or quasi-1D) treatment suffices. Moreover, the consequent simplifications—primarily that the flow *directions* do not have to be solved for—make the 1D equations especially useful for decoding the physics, and for elucidating the nature and form of the boundary conditions. It is for these reasons also that the 1D equations have been emphasized in the literature and will be focused on in this review. The geometry is as shown in Fig. 4.1.1 for a single barrier of width d , with bias V applied on the right electrode, and with forward (reverse) electrons flowing with (against) the bias. The treatment of hole tunneling is entirely analogous throughout, and so for simplicity will be ignored. We further specialize to steady state (and as noted before neglect space charge in the barrier), in which case the 1D versions of (4.1.1a), (4.1.1b) and (4.1.1c) for the forward and reverse electrons admit obvious integrals:

$$nv_n = J_n, \quad uv_u = J_u, \quad \psi = Vx/d \quad (4.1.2a)$$

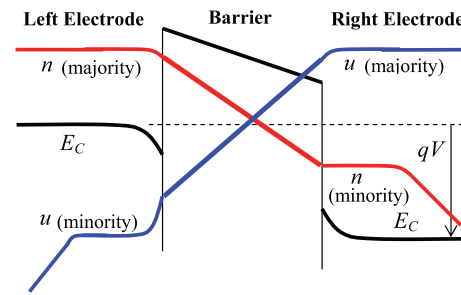


Fig. 4.1.1 Schematic depicting the macroscopic interpretation of tunneling

$$\begin{aligned} \psi - \phi_n^{DD} + \frac{2}{s}(b_n s_{,x})_{,x} - \frac{m_n}{2q}v_n^2 &= \Psi_n \\ \psi - \phi_u^{DD} + \frac{2}{z}(b_u z_{,x})_{,x} - \frac{m_u}{2q}v_u^2 &= \Psi_u \end{aligned} \quad (4.1.2b)$$

where J_n , J_u , Ψ_n and Ψ_u are integration constants, and $u \equiv z^2$ is the density of the reverse-flowing electrons.²³ As depicted in Fig. 4.1.1, each electron population (labeled n or u) is associated with a particular electrode and, as described by the foregoing equations, evanescence consists of the “leaking” of these carriers into the barrier. If appreciable numbers of carriers manage to traverse the barrier (as revealed by the presence of significant carrier density at the opposite electrode), then their “capture” and “conversion” to the carrier type of the downstream electrode is what constitutes tunneling. At zero bias ($V = 0$), the forward and reverse current flows will balance and zero net current will flow; as bias is applied the forward current grows quickly and the reverse current soon becomes negligible.

From a mathematical/computational standpoint, DGT theory differs from its classical analog (i.e., ballistic transport theory) in that with the latter the hyperbolic character of the left sides of (4.1.1b) is the main source of complexity. In DGT theory the high-order derivatives of the DG terms instead tend to dominate and this preserves much of the elliptic/parabolic character exhibited by the DGC equations (especially in 1D). Qualitatively, the DG terms in (4.1.2b) force strong exponential {“tunneling”} decays in the concentrations away from the emitting electrodes, and since (4.1.2a)₁ and (4.1.2a)₂ imply inverse relationships between density and velocity, the macroscopic kinetic energy terms in (4.1.2b) will be appreciable only where the densities are smallest. This generally occurs in the immediate vicinity of the downstream electrode, and thus throughout almost the entire barrier one can employ a “quasistatic” treatment in which inertial effects are entirely ignored. From a numerical

²³The choice of the variable names u and z for the reverse electron gas is a mnemonic introduced in [32] with these names being the “reverse” of n and s , respectively.

standpoint this fact together with the thinness of the barriers makes the DGT equations relatively simple at least in one dimension. Using a “Slotboom” approach as discussed in Sect. 3.1 [25], the author has had little trouble in solving these equations using either his own finite-difference code or with the finite-element method as implemented in the Comsol package see footnote 13.

4.2 Tunneling boundary conditions

As noted in Sect. 4.1, the boundary conditions of DGT theory differ from those of DGC theory because the directionality of the flows imposes a distinction between “upstream” and “downstream” boundaries (see Fig. 4.1.1). The upstream conditions are taken to be simple continuity conditions to reflect mathematically the idea that each tunneling carrier population is “part of” the carrier population in the upstream electrode. In contrast, the downstream conditions are both less clear and more complicated because they must quantify the capture of carriers that manage to traverse the barrier.

To aid in developing forms for the downstream conditions we focus first on the forward electron population that dominates the current except near zero bias. Assuming that the voltage drops in the electrodes are small, the electrons traversing in the forward direction will pick up energy equal to the entire applied voltage. Further assuming that all of these electrons are captured (rather than returning to the emitting electrode) then the average velocity of the electrons reaching the downstream electrode will be

$$v_n(x=d) \cong \gamma_n^{ideal}(ballistic) \equiv \sqrt{2qV/m_n} \quad \text{or} \quad J_n = \gamma_n s^2(x=d) \quad (4.2.1)$$

where the parameter $\gamma_n \equiv \gamma_n^{ideal}(ballistic)$ has been called the **tunneling recombination velocity** (TRV) [41]. Next, we assume that inside the downstream electrode the evanescent carriers that have arrived from upstream form an energetic “minority” carrier population n that merges with the native “majority” carriers of that electrode (of density u) at a rate determined by a thermalization lifetime τ_n . The continuity equations *inside* the downstream electrode then take the form

$$(nv_n)_{,x} = -n/\tau_n \quad \text{and} \quad (uv_u)_{,x} = u/\tau_n \quad (4.2.2)$$

Now, if we assume that v_n is slowly varying across the interface, then combining (4.2.2)₁ with (4.2.1) yields

$$s_{,x}(x=d) = -\frac{s(x=d)}{2\gamma_n\tau_n} \quad (4.2.3)$$

Equations (4.2.1)₂ and (4.2.3) were referred to as **generalized TRV conditions** in [44]. They contain two parameters—namely γ_n and τ_n —with an expression for the former being

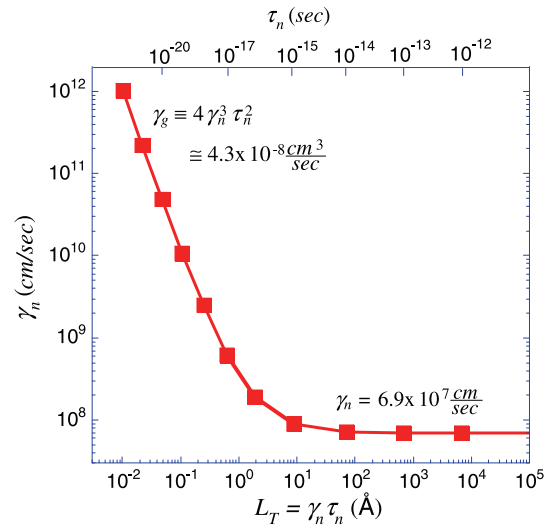


Fig. 4.2.1 Effect of scattering on the parameters defining ideal capture of electrons by the downstream electrode

given in (4.2.1)₁. With respect to the thermalization lifetime, it was shown in [44] that it is usually sufficient to consider only the limiting cases of τ_n being either “large” (elastic capture) or “small” (inelastic capture). The TRV conditions in these limits are readily shown to take the forms (that first appeared in [41]):

$$\text{TRV1: } s_{,x}(x=d) = 0 \quad \text{and} \quad J_n = \gamma_n s^2(x=d) \quad (\text{elastic capture, } \tau_n \rightarrow \infty) \quad (4.2.4a)$$

$$\text{TRV2: } s(x=d) = 0 \quad \text{and} \quad J_n = \gamma_g s_{,x}^2(x=d) \quad (\text{inelastic capture, } \tau_n \rightarrow 0) \quad (4.2.4b)$$

where the parameter γ_g is discussed further below. Although not fully explored, there is evidence that the TRV1 conditions are more appropriate when the electrodes are semiconductors [44], whereas the TRV2 conditions seem to be better for metal electrodes [45]. In any case, it is to be emphasized that, within the assumptions made (and especially (4.2.1)₁), the TRV1 and TRV2 conditions contain no free parameters.

In order to investigate the behavior between the TRV1 and TRV2 limits, we observe that scattering in the downstream electrode cannot deliver more current than is obtained in the elastic limit. On this basis, we then consider a situation that includes scattering (with a given τ_n) and obtain a value for $\gamma_n \equiv \gamma_n^{ideal}(scattering)$ by curve-fitting the DG current in the no-scattering limit (for a particular semiconductor-insulator-semiconductor diode under a given bias) using quasistatic DG theory [44] with (12). The results, plotted as γ_n versus a thermalization length $L_T \equiv \gamma_n \tau_n$, appear in Fig. 4.2.1. When the scattering is weak we see that γ_n is essentially constant (independent of τ_n); again this is the “elastic capture” regime governed by the TRV1 conditions

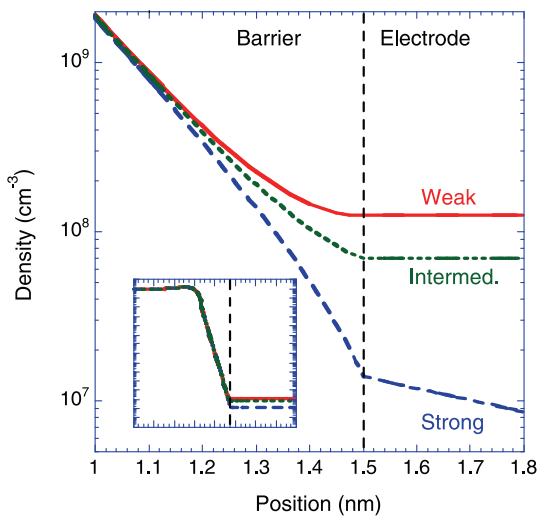


Fig. 4.2.2 The effect of scattering in the downstream electrode on the right-going electron density profile

in (4.2.4a). In the case of strong scattering, the asymptotic ideal obtains when

$$\gamma_n \cong \gamma_n^{\text{ideal}}(\text{scattering}) \equiv a_n \tau_n^{-2/3} \quad (4.2.5)$$

where a_n is a voltage-dependent constant that depends on $\gamma_n^{\text{ideal}}(\text{ballistic})$ through the curve-fitting, and is also related to the constant in (4.2.4b)₂ by $\gamma_g = 4a_n^3$. Figure 4.2.1 shows that the TRV1 and TRV2 limits actually cover almost the entire range, and that only in the narrow range of $L_T \sim 2\text{--}10\text{\AA}$ is an intermediate scattering regime seen. In this regime the conditions in (4.2.4a) and (4.2.4b) are not good approximations, and one instead needs to use the generalized conditions in (4.2.1)₂ and (4.2.3) which depend on τ_n explicitly. For a particular SIS diode, a few solution profiles for situations of ideal weak, strong and intermediate scattering as defined by Fig. 4.2.1 are shown in Fig. 4.2.2. The inset displays the entire profile, while the main plot is a close-up in the vicinity of the downstream electrode. As the inset shows, the solutions are indeed largely unaffected by the downstream BC and its representation of the scattering, and it is only in the immediate vicinity of the downstream electrode that the solutions are altered by the details of the electron capture process. The observed variations in density (as well as the concomitant rise in γ_n according to (4.2.5)) suggest that the scattering is causing the electron gas to accelerate as it gets very close the contact (within about 1–2 Å). It is likely that this is an artifact of the low-order equations attempting to represent the extremely rapid conversion and disappearance of the tunneling electrons as they enter the downstream electrode.

With respect to the reverse current, the considerations are much the same but there is one main difference in SIS devices. Specifically, as depicted in Fig. 4.2.3, for electrons

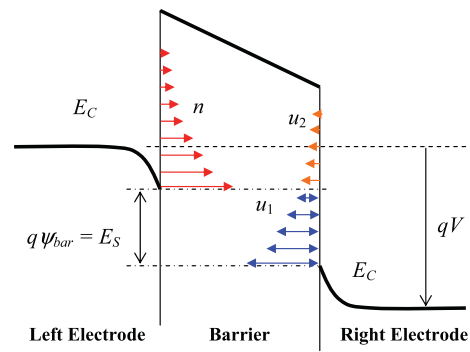


Fig. 4.2.3 Schematic showing the forward and reverse tunneling populations and the bandgap blocking effect that acts on the reverse population u_1

in the population u_1 , there will be a “bandgap blocking effect” [32] wherein the lack of final states prevents any capture/tunneling from occurring. As shown, evanescence of these carriers will still occur, and in this way they could still influence behavior through the electrostatics, however, this is unlikely ever to be important. For understanding reverse tunneling we can therefore focus entirely on the population u_2 in Fig. 4.2.3. Since the equation of state is independent of the velocity, the chemical potential of this gas can be estimated using statistical mechanics and the upstream density will be given by

$$u_2 = \int_{E_C + E_S}^{\infty} f(E)g(E)dE \cong \frac{2u}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}, \frac{E_S}{k_B T}\right) \quad (4.2.6)$$

where $f(E) = \frac{1}{1 + \exp[(E - E_i - q\phi_u^0)/k_B T]}$

where E_S is the “splitting energy” defined in Fig. 4.2.3, $g(E)$ and ϕ_u^0 are respectively the density of states and chemical potential in the right electrode, and the usual approximation of making the upper limit of the integral infinity has been instituted. The approximation in (4.2.6) is that appropriate for parabolic bands and Maxwell-Boltzmann statistics where Γ is the incomplete gamma function. Not surprisingly, u_2 approaches $u \exp(-E_S/k_B T)$ for $E_S \gg k_B T$. Lastly, arguments like those given in relation to (4.2.1)₁ provide an estimate of the capture velocity at the left electrode of $\gamma_u = \sqrt{[k_B T \ln(u_2/u) + qV]/m_u}$.

4.3 Elastic tunneling in 1D: theoretical comparisons

In this section we analyze 1D semiconductor-insulator-semiconductor (SIS) diodes with heavily doped n-type contacts as depicted in Fig. 4.1.1 using DGT theory and assess its accuracy and range by comparing its results with those of quantum mechanics. As in Sect. 3, such comparisons between theories (as opposed to comparisons with experiment [32]) are especially instructive because of the greater knowledge and control one has over physical content. In the

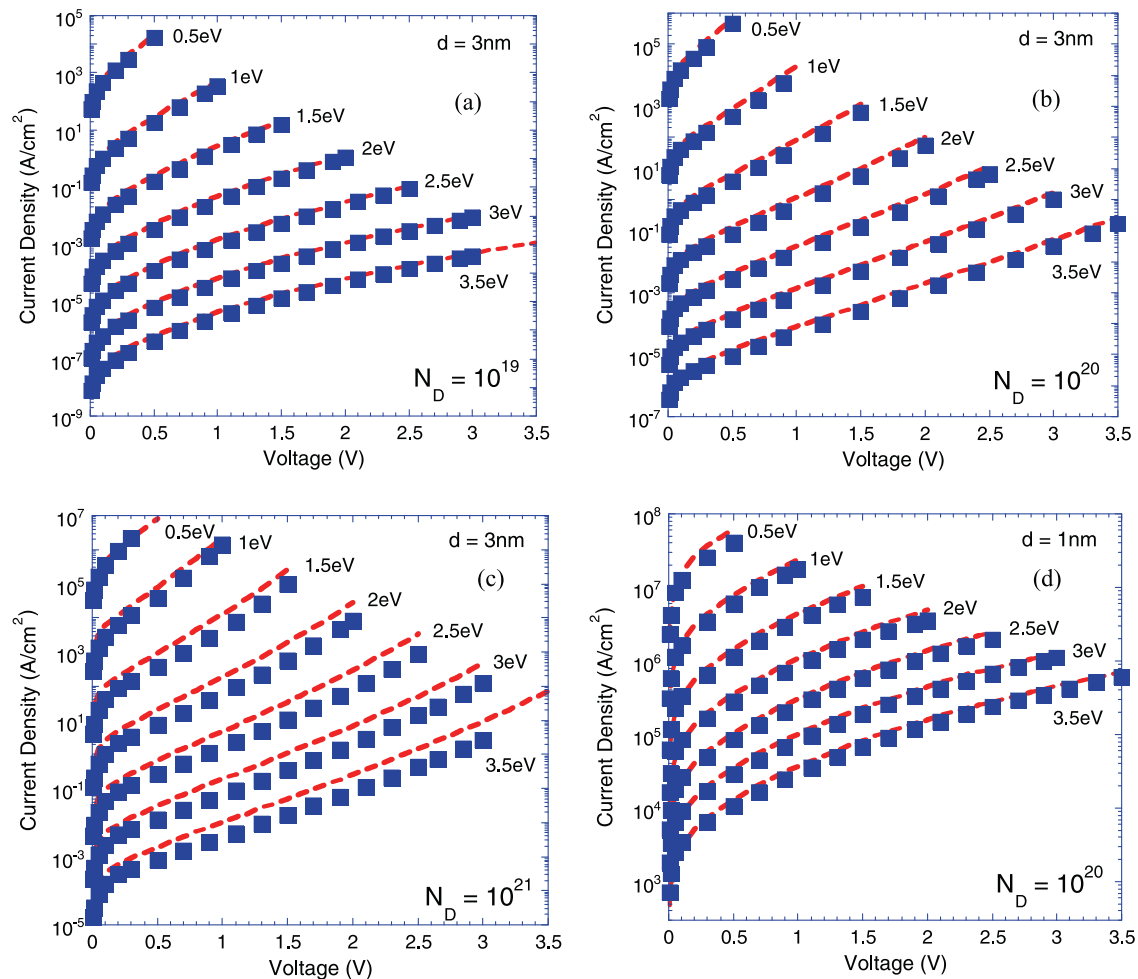


Fig. 4.3.1 Current-voltage characteristics of SIS diodes as computed by NEGF (squares) and DG (dashed lines) with varying barrier heights for a 3 nm barrier with electrode dopings of (a) 10^{19} cm^{-3} , (b) 10^{20} cm^{-3} and (c) 10^{21} cm^{-3} , and (d) a 1 nm barrier with electrode doping of 10^{20} cm^{-3}

past, such studies were used to examine the DGT treatment of tunneling from both metal [45] and semiconductor [44] electrodes with the latter being emphasized here.

The DG solutions are obtained by solving the DGT equations of Sect. 4.1 with $r_n = 1$ inside the barrier and using the tunneling BCs of Sect. 4.2. For the transport in the semiconductor electrodes it is assumed for convenience that strong scattering conditions prevail so that the appropriate equations are those of DGC theory with r_n usually taken to be 3 (see Sect. 3). The analogous quantum mechanical analyses are performed using the NEGF approach assuming elastic tunneling. Of course for the comparisons to be most meaningful, to the extent possible the two methods need to examine the same physical situation, a potentially tricky proposition given that one theory is microscopic and the other macroscopic. However, our matching is much aided by the fact that the quantum mechanical treatment is itself not fully microscopic in that it contains a number of macroscopic elements. Thus the NEGF invokes the effective-mass approximation with both theories using the same electron ef-

fective mass ($0.328m_0$) as well as identical barrier heights, uniform (jellium) doping densities and dielectric properties. To match DGT theory, the NEGF calculations also include the Hartree electrostatics. Lastly, regarding scattering both treatments are scattering-free within the barrier. That the DG calculations include scattering in the upstream electrode whereas the NEGF does not, is not expected to have much impact because of the heavy doping. At the downstream electrode the DG simulations ignore scattering by using the TRV1 condition in (4.2.4a).

As a first set of DG/NEGF comparisons, in Figs. 4.3.1a–4.3.1d we plot the current density versus voltage (J – V) for SIS diodes with various barrier heights, electrode doping the levels and barrier thicknesses. In general, the DG predictions are in excellent agreement with the NEGF calculations, and especially when one considers the exponential nature of the tunneling currents. The fact that these plots involve no curve fitting means these agreements are strong evidence in favor of DGT theory’s macroscopic representation of the physics.

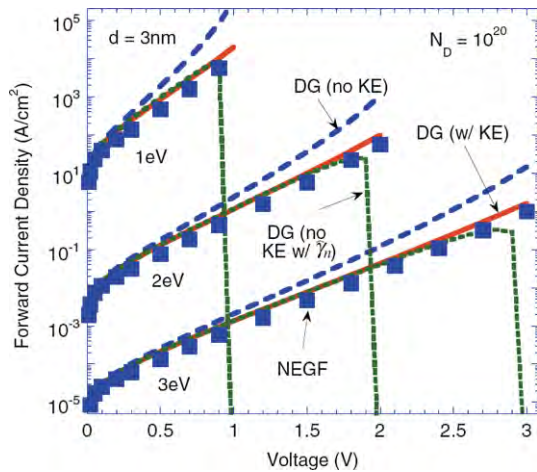


Fig. 4.3.2 Forward current for SIS diodes with three different barrier heights as computed by DG theory with and without the kinetic energy included. Also shown is a simple correction to the calculation without kinetic energy that leads to good agreement over most of the range

As emphasized earlier, one key aspect of DGT theory is its assertion that the kinetic energy of the tunneling electron gas can be split into macroscopic and microscopic contributions. To test this assumption directly, in Fig. 4.3.2 we compare DG calculations for a few SIS diodes as calculated with (solid curve) and without (dashed curve) the kinetic energy term included. Also shown in the figure are the NEGF results for these diodes (squares). Clearly, the kinetic energy term is needed for the DG simulations to be most accurate, and not surprisingly this is particularly true at higher bias. Thus DG theory's fundamental splitting of the kinetic energy does indeed appear to be valid.²⁴ Finally, that the kinetic energy is important only in the immediate vicinity of the downstream electrode (as noted in Sect. 4.2) makes it possible to subsume its effect into the TRV1 condition. In particular, it was shown in [44] that if one multiplies γ_n by the factor $(1 - V/\phi_B)$, one can achieve a good representation (see Fig. 4.3.2, dotted curve) over most of the bias range using a “quasistatic” treatment in which the kinetic energy term is not included explicitly in the differential equation (4.1.2b).

Although the agreements between the J - V characteristics calculated by DGT theory and NEGF in Figs. 4.3.1a–4.3.1d are quite good, discrepancies do exist and are seen to grow with increasing electrode doping and barrier height. To understand the DG errors²⁵ better we next compare field solutions. A first set of comparisons is shown in Figs. 4.3.3a

²⁴This conclusion has implications for transient behavior as well but fidelity in this regard would be hard to demonstrate.

²⁵These differences between the DG and NEGF simulations are regarded as DG errors. However, since the two theories differ in physical content, it is conceivable that the DG results could actually correspond better with experiment much as in the quantum confinement situation discussed in Fig. 3.2.7.

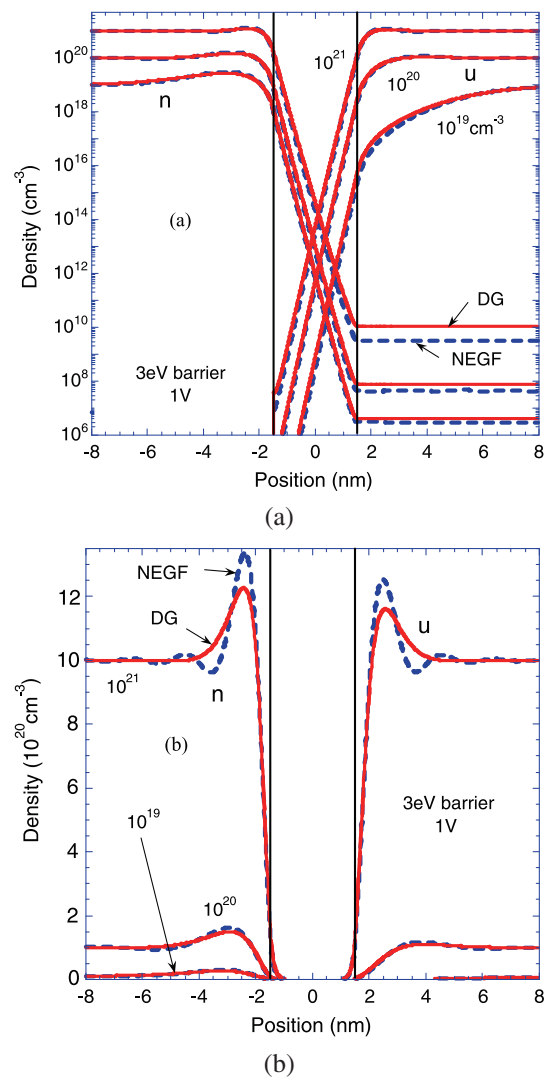


Fig. 4.3.3 (a) Semi-log and (b) linear density profiles as calculated by NEGF (dashed) and DG (solid) profiles in an SIS diode with three different electrode doping concentrations

and 4.3.3b where we plot the electron density profiles as computed by NEGF and DG for SIS diodes with $\phi_B = 3$ eV, $d = 3$ nm, $V = 1$ V, and $N_D = 10^{19}$, 10^{20} or 10^{21} cm⁻³. Generally the picture is one of superb agreement and, as before, this is especially so when one considers the extremely wide variation in densities involved. The most important errors are the small differences in the minority carrier density (n) in the right electrode seen in Fig. 4.3.3a since these relate directly to the current via (4.2.4a). In accord with Fig. 4.3.1a–4.3.1c, these errors grow with increasing doping. A second DG error, most evident in Fig. 4.3.3b and again amplified by doping, is in the density profiles inside both electrodes. These screening errors were first noted in the context of metals where they are especially large [45] and can be traced to DG theory's neglect of higher-order gradient effects [44, 45]. Fundamentally this error is not a tunneling error but a confinement error, and indeed has al-

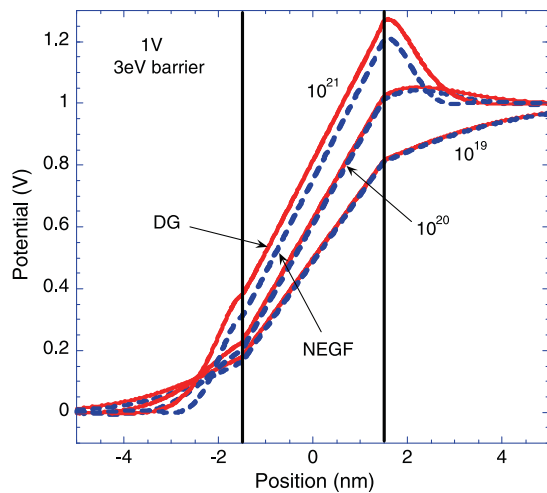


Fig. 4.3.4 Electric potential profiles across various SIS diodes and showing their “image charge” barrier-lowering effect

ready been seen in that connection in Sect. 3.2 and especially Fig. 3.2.6. As before, not only are the profile shapes incorrect but there is also a complete absence of the Friedel oscillations that characterize the quantum mechanical simulations [31]. In the case of tunneling the most important consequence of these screening errors is that they actually cause the errors in the downstream density in Fig. 4.3.3a and hence produce the errors in the calculated currents in Fig. 4.3.1. To see this, we note that the dominant electrostatic effect at these concentrations is a *barrier-lowering* commonly referred to as an “image charge effect” that is associated with the proximity of the two electrodes [45]. A plot of the barrier lowering for the cases of Figs. 4.3.3a–4.3.3b as computed by NEGF and DG is shown in Fig. 4.3.4. Qualitatively, it is clear that DG theory is capturing much of the physics associated with the coupled electrostatics. However, quantitatively it is also evident that there is a growing error in the barrier-lowering with increased doping. At a doping of 10^{21} cm^{-3} , the error is about 69 mV. (Because of the differences in density, there is also an error in the chemical potential but this turns out to be only about 12 mV, so the electrostatic effect is dominant.) As dictated by (4.1.1a), (4.1.1b) and (4.1.1c), the barrier height error translates directly into an error in the exponent that sets the evanescent decay inside the barrier, that is discernable as an error in the *slopes* in Fig. 4.3.3a, and that leads to the observed errors in the downstream densities. This then demonstrates that the main errors in Figs. 4.3.1a–4.3.1d are not the result of a flaw in the DGT treatment of quantum mechanical tunneling, but rather stem from inadequacies in DG theory’s representation of quantum confinement as already discussed in Sect. 3.

One final error in the forward-density is a discontinuity in n at the surface of the downstream (right) electrode in the DG-NEGF comparison at $V = 0.2 \text{ V}$ in Fig. 4.3.5. This discontinuity originates in the assumption of the weak-scattering electrode (see Sect. 4.2) wherein the transport is

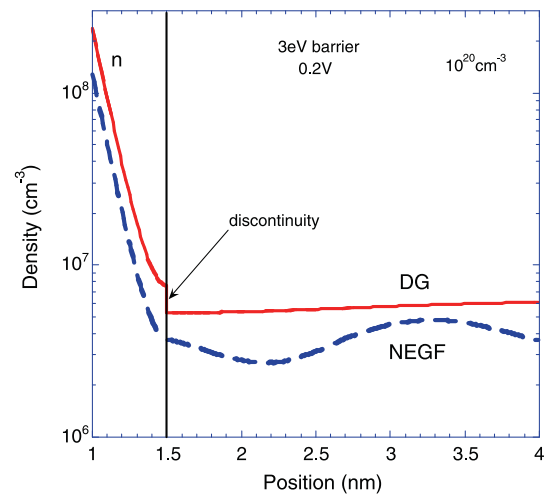


Fig. 4.3.5 Close-up of a portion of Fig. 4.3.3a near the surface of the right electrode

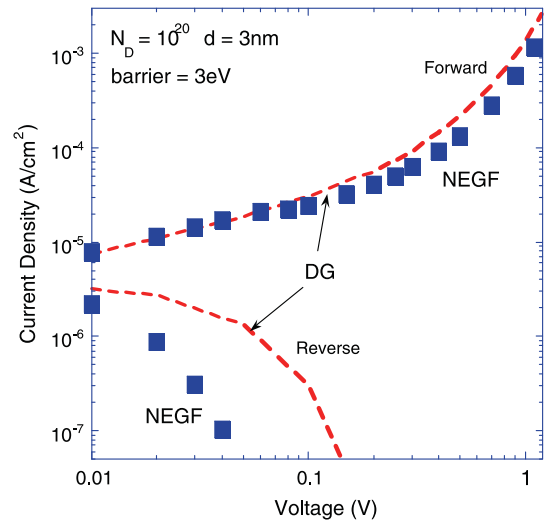
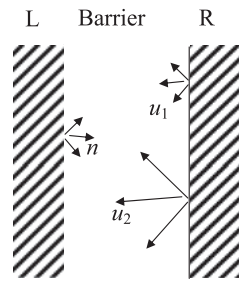


Fig. 4.3.6 Comparison of forward and reverse J - V characteristics as computed by NEGF and DG for a particular case

taken to be purely ballistic. The small size of the discontinuity suggests that this assumption is quite good. Moreover, as the bias increases from this very small value, the discontinuity becomes even smaller (data not shown) because of the larger kinetic energy. Lastly, we note from Fig. 4.3.5 that the NEGF result also exhibits some density oscillations in the downstream electrode that are not present in the DG simulation.

As noted earlier, the current contribution of the reverse-density u is much less important than that of the forward-density in that it plays a significant role only at very low biases. This may be seen in the DG-NEGF comparison shown in Fig. 4.3.6 where we plot the forward and reverse currents versus applied bias. The error in the reverse current (as well as the accuracy of the forward current) is apparent, as is the fact that this error decreases as the bias is reduced. As a result, the DG-calculated reverse current is fortuitously

Fig. 4.3.7 Illustration of the effect of transverse momentum on tunneling of u_2



becoming most accurate just at the biases where it is most important (below a few $k_B T$). Indeed, this error vanishes at zero bias when the forward and reverse currents come into balance yielding zero total current. The origin of this behavior is not entirely certain, but it is most likely due to the significant number of electrons in the important population u_2 (see Fig. 4.2.3) that have high transverse momentum. As depicted schematically in Fig. 4.3.7, the anisotropy will cause such electrons to have a reduced average tunneling probability and will thereby decrease the total reverse current. This explanation is consistent with Fig. 4.3.6 and DG theory's growing over-estimation of the reverse current as the applied bias increases.

4.4 Elastic tunneling in 1D: phenomenology

The primary error made by DG theory in modeling tunneling originates, as discussed in connection with Fig. 4.3.4, in inaccuracies in DG theory's representation of the high-density screening layers inside the semiconductor electrodes. This flaw of DG theory is well known though not fully understood, having been noted in the contexts of both quantum confinement in Sect. 3 and also in field emission from metals in [45]. A physically well-motivated approach for improving the representation would be to develop a higher-order DG theory with the hope that including higher-order gradient terms would provide significant improvements. This physics-based possibility is not pursued here however. Other ways of refining the description are phenomenological in nature, essentially using DG theory as the mathematical framework and treating one or more of its coefficients as fitting parameters. Although in past work on tunneling we did such fitting using γ_n and γ_g [41], given our identification of the source of the error as being in the treatment of the quantum confinement, it seems better to use the surprisingly accurate phenomenology of Sect. 3 with the quantity r_n in (2.5.7a)₂ as the fitting parameter.

With this as motivation, we develop a phenomenological treatment of the barrier-tunneling problem of the previous section by using r_n as a fitting function as done in [44]. Demonstrating the efficacy of this approach, if we allow r_n to vary with density as shown in Fig. 4.4.1, then the comparisons of Figs. 4.3.1a and 4.3.1c are now as shown in Figs. 4.4.2a and 4.4.2b. The improved agreement is obvious.

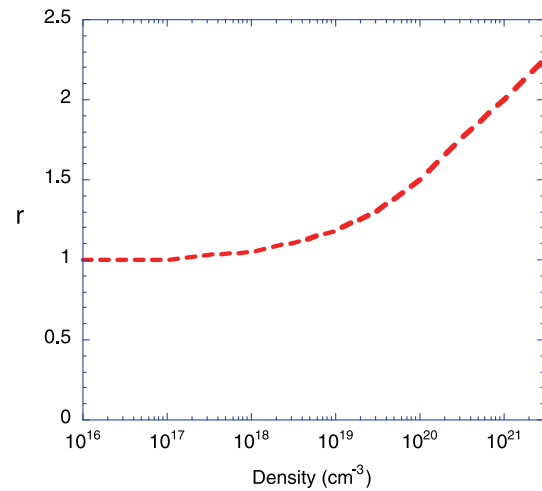


Fig. 4.4.1 Assumed variation of r_n as a function of density in order to get the curvefits displayed in Figs. 4.4.2a, 4.4.2b

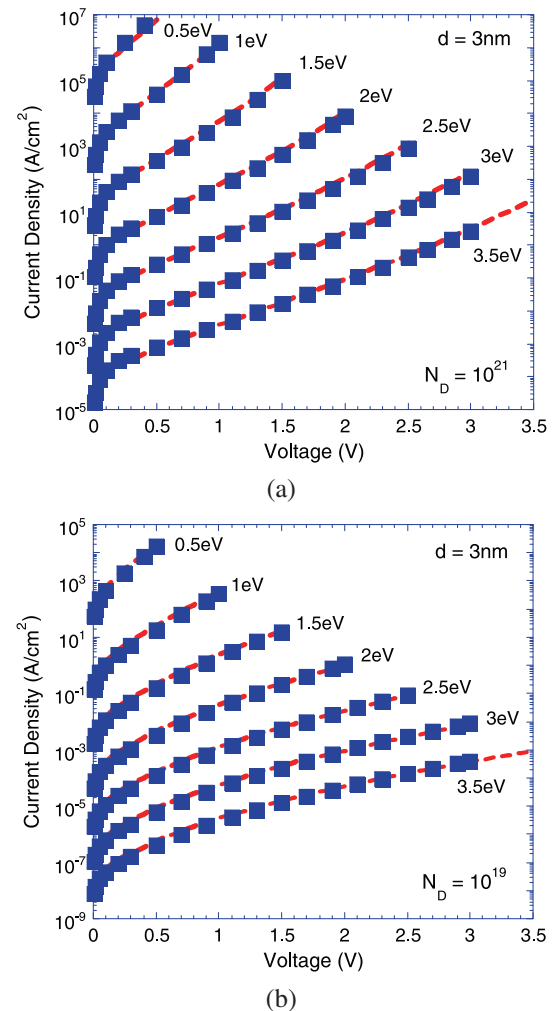


Fig. 4.4.2 Identical plots as shown in Figs. 4.3.1a and 4.3.1c with electrode dopings of (a) 10^{19} and (b) 10^{21} cm^{-3} except that the DG calculations were obtained with the parameter r_n selected to vary with density as shown in Fig. 4.4.1

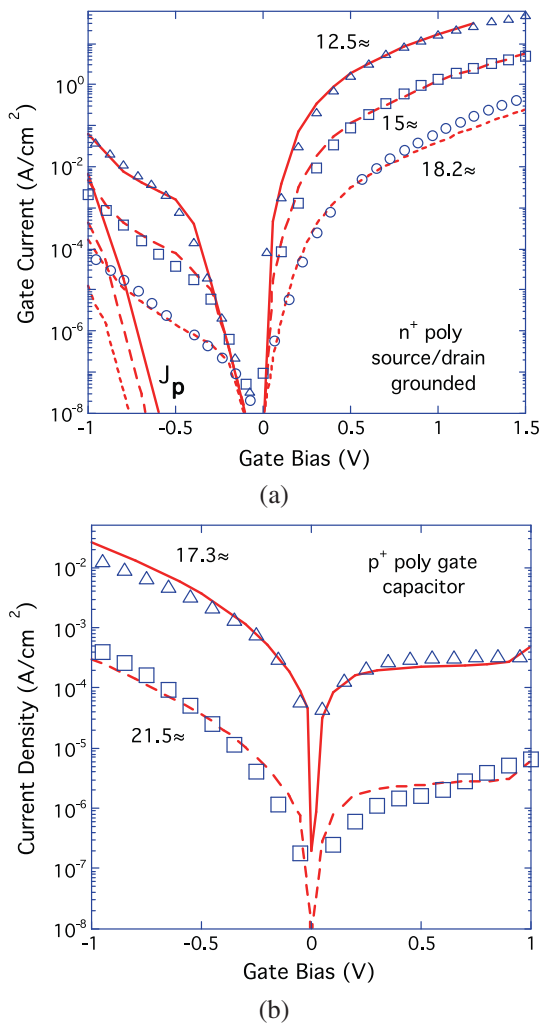


Fig. 4.5.1 Comparison of DGT simulations with experimental gate current characteristics from [32] with various gate oxide thicknesses and for (a) n^+ poly gates and (b) p^+ poly gates

Evidence that this is merely curve-fitting is to be found in the fact that the calculated density profiles obtained in this way (not shown) are not noticeably better than those displayed in Fig. 4.3.4. Thus, curve-fits of this sort should be viewed as no more than mathematical representations that may be of value for engineering.

One final form of tunneling phenomenology that has already been discussed (Sect. 3.5) is that of using DGC theory. As noted in the earlier section, this approach is especially effective in complicated situations such as that of source-drain tunneling in a short-channel FET. A DGT approach to this same problem seems unworkable because it would require multiple tunneling populations as well as a need to define precisely where the tunneling is occurring.

4.5 Elastic tunneling in 1D: experimental comparisons

As discussed in Sect. 3, comparisons of theory with experiment are generally harder to make. From the experimental

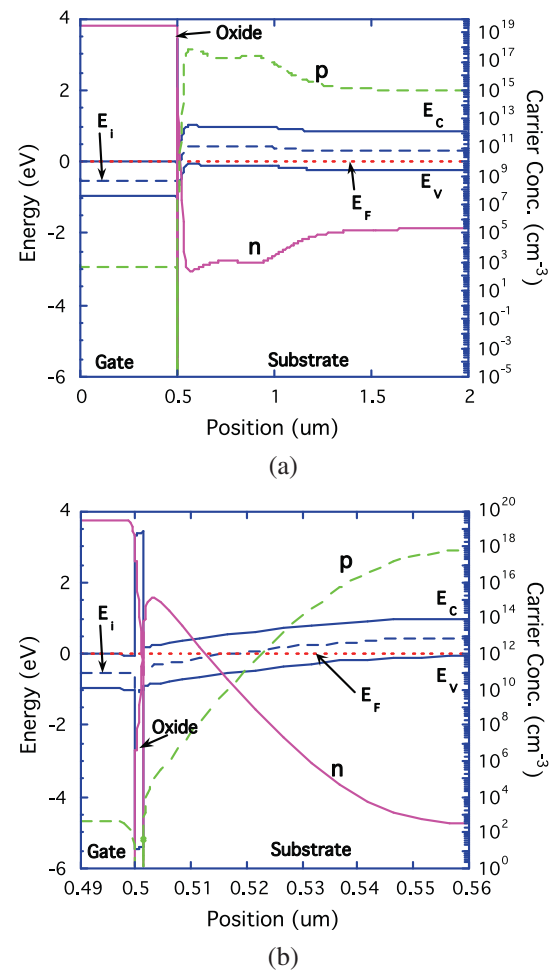


Fig. 4.5.2 Energy band diagram and carrier profiles in the DGT approximation for an n^+ -poly MOSFET ($t_{ox} = 15 \text{ \AA}$) with zero gate voltage and the source floating. (a) Shows the full device while (b) shows a close-up of the barrier region

side a main difficulty is insufficient knowledge, e.g., about doping profiles, surface roughness, or defects. And theory can be deceptive if any curve-fitting of experiment is done, since this can easily hide modeling deficiencies. Nevertheless, experiment is reality, and so it is always the ultimate test. For DGT theory the most meaningful comparisons with experiment to date appeared in [32], and a few results from that work are briefly reviewed next.

The experiments of [32] were measurements of gate current in carefully characterized MOS capacitors and gated diodes with ultra-thin oxides. Basic comparisons of DGT theory and experiment for n^+ and p^+ poly gate devices with several different gate oxide thicknesses are shown in Figs. 4.5.1a and 4.5.1b, respectively. Both electron and hole tunneling are included and one curve-fit is involved, namely of an “effective” tunneling area that is a fixed multiplicative factor in all the curves. The qualitative shapes of the results are determined by a combination of the tunneling and band-bending in the channel. Clearly, the agreements are excel-

lent. This is additional evidence of the validity of the DGT description, and is also an illustration of the method's utility when applied to real device situations. A further demonstration of the latter point is in Fig. 4.5.2a and the close-up in Fig. 4.5.2b which show the band diagram and density profile across an MOS capacitor. What is most interesting about these plots is that they exhibit the ability of the single set of differential equations of DG theory to capture both the fine details of the barrier tunneling that occur on an Ångström-scale and the ordinary device physics that plays out over micron scales.

5 Final remarks

The singular fact of the electronics revolution is the extraordinary half-century of exponential progress that has taken place in accord with Moore's Law. However, notwithstanding a long history of erroneous predictions,²⁶ it seems safe to say that the “endgame” has now finally begun. In striving for further advances, this concluding epoch—which will surely define what the buzz word “nanoelectronics” really means—will clearly be characterized by an increasingly varied set of physical phenomena, materials, processes, device geometries and circuit architectures. In this expanding universe, the technical obstacles and the accompanying economic costs will continue to escalate,²⁷ and among other things, this will open up many new challenges and opportunities for modeling and simulation. And with respect to the topic of this review, the physical phenomena of quantum confinement and quantum tunneling will undoubtedly grow ever more ubiquitous, and so DG theory is sure to remain relevant.

As presented in this review, DG theory possesses a substantial capacity for the analysis of current flows in semiconductor devices when quantum confinement and tunneling effects are important. In many cases, this capability comes from the DG description being physically well founded, but just as often the approach's vitality stems from its surprising potency as a phenomenology. These demonstrated powers are likely to make the DG approach a valuable tool for future device engineering. Beyond this, it is important to recognize that DG theory remains a work in progress, and like all classical field theories, it constitutes a flexible framework in which broader generalizations and further extensions can be investigated in a consistent fashion.

The key to this potentiality are the material response functions discussed in Sect. 2.2. (Another as-yet-unexplored avenue for generalization is the possibility of higher-order versions of the theory that include dependences on the second-gradients, etc.) Some of the areas for which DG theory might well prove useful (including some for which initial studies have already been conducted) are multi-dimensional tunneling [47], Fowler-Nordheim tunneling [41], reverse-bias Schottky barrier tunneling, Zener tunneling, and field emission from metals [45].

Given the amazing longevity of DD theory, perhaps even more important reasons that DG theory will remain utile are its familiarity, robustness, and efficiency as compared with microscopic alternatives.²⁸ The DG approach is of course familiar both as a straightforward generalization of DD theory and mathematically as a coupled set of partial differential equations. With regards to robustness, continuum approaches are justly renowned for this attribute [1], and DG theory is especially resilient because of the many sources of smoothing inherent in electronic devices, e.g., quantum mechanical “smearing” of light-mass electrons/holes, temporal averaging, width averaging, etc. Finally, the efficiency of the DG approach originates in part from the ease with which its PDEs can be solved numerically, a benefit of years of research on PDE methods and the associated software infrastructure. A second source of efficiency is the fact that DG theory can readily provide unified treatments of “complex” devices, e.g., of a large multi-dimensional classical device with an embedded tunnel barrier. By contrast, microscopic methods tend to be far inferior, especially in often being extremely intensive computationally (particularly when scattering and the full self-consistent physics is included), and notoriously ill-suited to marrying different levels of treatment (e.g., Boltzmann theory in one region with NEGF in another). Another major weakness of microscopic treatments is their dependence on microscopic assumptions (e.g., scattering cross-sections at high energy) whose necessity, validity and range are often hard to ascertain.

Despite DG theory's (and DG phenomenology's) power and flexibility, one should not lose sight of the fact that the approach is still an approximate one. For the description of quantum effects, this has two major consequences. First, as we saw from DG theory's lack of Friedel oscillations (Sects. 3.2 and 4.3), the theory does not describe interference phenomena or, more generally, any effect in which phase plays a determinative role. Most critically, this means that there are many potential devices like those of molecular electronics for which only a full quantum mechanical

²⁶Interestingly, such predictions actually antedate Moore's Law by a few years with the earliest public expression known to this author being in 1961 [49].

²⁷One probable consequence of the economics is a continued narrowing of the number of chipmakers who can afford to remain in the “game”.

²⁸Over the years, academic researchers have often suggested that the utility of continuum approaches was coming to an end, and that they would soon have to be replaced by microscopic methods even for device engineering. Curiously paralleling the incorrect technology forecasts of the past, all such predictions have to date proven premature.

description will suffice. The second area of weakness comes not from a failure of the DG approximation *per se*, but rather from a failure of its underlying continuum approximation. An example involving quantum transport is seen in tunneling situations with non-abrupt barriers where, in effect, one has a different barrier at each energy so that a representation in terms of a single (or a few) well-defined tunneling population does not seem tenable. How serious this latter error is remains to be seen, but the success of a simple DGC phenomenology in modeling source-drain tunneling (see Sect. 3.5) gives cause for optimism.

Given the comparative advantages and disadvantages of the macroscopic and microscopic approaches to quantum transport, it is best not to rely solely on one class of methods over the other. Instead they should be regarded as complementary with each being a potentially valuable weapon in the service of device engineers and designers. Which to use largely depends on the nature of the application and on one's motivation (e.g., device design using well-understood materials or predicting ultimate performance of some new material like graphene). As a tactical choice for engineering applications, this author's view is that if a continuum description is available, then it should "always" be used, with microscopic theory being used solely (if at all) as a means of estimating the macroscopic and/or phenomenological coefficients such as the DG effective mass. Of course if a continuum description is untenable, then the only available theoretical route would be a full quantum transport theory such as that provided by NEGF.

Finally, to close with a provocative supposition it may not be mere coincidence that macroscopic descriptions of electron transport are being pushed to their limits at just the same time that Moore's Law is coming to an end. The logic behind this assertion is that ultra-small devices without the "large numbers" needed to smooth out various sources of randomness (e.g., thermal, quantum, geometric or dopant fluctuations) and hence to ensure regular, predictable behavior will also tend to be ones that lack the well-defined averages upon which continuum descriptions depend. If this admittedly speculative claim is correct, then it could be that a *sine qua non* for any future nanoelectronics technology is that its electron transport be such that it is describable by macroscopic modeling methods! And if not, then the time may at long last have come when microscopic descriptions of electron transport and the requisite computational infrastructure must finally mature to the point of providing real-world engineering tools.

Acknowledgements The author thanks Professor Zhiping Yu of Tsinghua University for his encouragement over many years, as well as for his kind invitation to write this article. Thanks also go to Dr. Andrew Brown of the University of Glasgow for generously allowing the use of Figs. 3.5.5–3.5.7 from [42], and to Dr. Chagaan Baatar and the Office of Naval Research for funding support. Lastly, with respect and gratitude, the author again acknowledges his teacher, the late Professor

Harry F. Tiersten of the Rensselaer Polytechnic Institute, whose deep physical insight and uncompromising devotion to truth and clarity have long been an inspiration.

References

1. Truesdell, C.A., Toupin, R.: The classical field theories. In: Handbuch der Physik, vol. III/1, Springer, Berlin (1960)
2. Shockley, W.: Electrons and Holes in Semiconductors. Van Nostrand, London (1951)
3. van Roosbroeck, W.V.: Theory of flow of electrons and holes in Germanium and other semiconductors. Bell Syst. Tech. J. **29**, 560 (1950)
4. Maxwell, J.C.: On stresses in rarefied gases arising from inequalities of temperature. Philos. Trans. R. Soc. Lond. A **170**, 231 (1876)
5. Wigner, E.: On the quantum correction for thermal equilibrium. Phys. Rev. **40**, 749 (1932)
6. Bloch, F.: Bremsvermögen von Atomen mit mehreren Elektronen. Z. Phys. **81**, 363 (1933)
7. von Weizacker, C.: Zur Theorie der Kernmassen. Z. Phys. **96**, 431 (1935)
8. Ancona, M.G., Tiersten, H.F.: Macroscopic physics of the silicon inversion layer. Phys. Rev. B **35**, 7959 (1987)
9. Hohenberg, P.C., Kohn, W.: Inhomogeneous electron gas. Phys. Rev. **136**, B864 (1964)
10. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. Phys. Rev. Lett. **77**, 3865 (1996)
11. Wilson, C.L.: Hydrodynamic carrier transport in semiconductors with multiple band minima. IEEE Trans. Electron Devices **35**, 180 (1988)
12. Ancona, M.G.: Hydrodynamic models of semiconductor electron transport at high fields. VLSI Des. **3**, 101 (1995)
13. Selberherr, S.: Analysis and Simulation of Semiconductor Devices. Springer, Vienna (1984)
14. Bohm, D.: A suggested interpretation of the quantum theory in terms of hidden variables. Phys. Rev. **85**, 166 and 180 (1952)
15. A thorough development of DG equations from moment expansions of the Wigner-Boltzmann equation appears in Gardner C.L. The quantum hydrodynamic model for semiconductor devices. SIAM J. Appl. Math. **54**, 409 (1994)
16. Jungemann, C., Meinerzhagen, B.: Hierarchical Device Simulation. Springer, Vienna (2003)
17. See Perrot, F.: Gradient correction to the statistical electronic free energy at nonzero temperatures: application to equation-of-state calculations. Phys. Rev. A **20**, 586 (1979) and references therein. For derivations in a semiconductor context see M.G. Ancona, G.J. Iafrate, Quantum correction to the equation of state of an electron in a semiconductor. Phys. Rev. A **39**, 9536 (1989) and M.G. Ancona, Finite temperature, density gradient theory, Proc. Comput. Electron. Workshop, 151 (1992)
18. Mermin, N.D.: Thermal properties of inhomogeneous electron gas. Phys. Rev. **137**, 1441 (1965)
19. Ancona, M.G.: Density gradient theory analysis of electron distributions in heterostructures. Superlattices Microstruct. **7**, 119 (1990)
20. Messiah, A.: Quantum Mechanics, p. 222. North Holland, Amsterdam (1965)
21. Pinnau, R.: A review of the quantum drift-diffusion model. Transp. Theory Stat. Phys. **31**, 367 (2002)
22. de Falco, C., Jerome, J.W., Sacco, R.: Quantum-corrected drift-diffusion models: solution fixed point map and finite element approximations. J. Comput. Phys. **228**, 1770 (2009)

23. Ancona, M.G.: Asymptotic structure of the density-gradient theory of quantum transport. In: Proc. Workshop on Computational Electronics, (1990)
24. Uno, S., Abebe, H., Cumberbatch, E.: Analytical description of inversion-layer quantum effects using the density gradient model and singular perturbation theory. *Jpn. J. Appl. Phys.* **26**, 7648 (2007)
25. Slotboom, J.: Iterative scheme for 1- and 2-dimensional dc transistor simulation. *Electron. Lett.* **5**, 677 (1968)
26. Wettstein, A., Schenk, A., Fichtner, W.: Quantum device simulation with the density-gradient model on unstructured grids. *IEEE Trans. Electron Devices* **48**, 279 (2001)
27. Ancona, M.G.: Finite-difference schemes for the density-gradient equations. *J. Comput. Electron.* **1**, 435 (2002)
28. Odanaka, S.: Multidimensional discretization of the stationary quantum drift-diffusion model for ultrasmall MOSFET structures. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **23**, 837 (2004)
29. Wettstein, A., Penzin, O., Lyumkis, E.: Integration of the density gradient model into a general purpose device simulator. *VLSI Des.* **15**, 751 (2002)
30. Ancona, M.G.: Equations of state for silicon inversion layers. *IEEE Trans. Electron Devices* **47**, 1449 (2000)
31. Ashcroft, N.W., Mermin, N.D.: *Solid State Physics*. Holt-Winston, New York (1976)
32. Ancona, M.G., Yu, Z., Dutton, R.W., Voorde Vande, P.J., Cao, M., Vook, D.: Density-gradient analysis of MOS tunneling. *IEEE Trans. Electron Devices* **47**, 1449 (2000)
33. See, e.g., Spinelli, A., Benvenuti, A., Pacelli, A.: Self-consistent 2-D model for quantum effects in n-MOS transistors. *IEEE Trans. Elect. Dev.* **45**, 1342 (1998). A comparison of quantum drift-diffusion and DGC theory appeared in Baccarani, G., Gnani, E., Gnudi, A., Reggiani, S., Rudan, M.: Theoretical foundations of the quantum drift-diffusion and density-gradient models. *Solid-State Electron.* **52**, 526 (2008)
34. Zhou, J.-R., Ferry, D.K.: Simulation of ultra-small GaAs MES-FET using quantum moment equations. *IEEE Trans. Electron Devices* **39**, 473 (1992)
35. Zhou, J.-R., Ferry, D.K.: *IEEE Trans. Electron Devices* **39**, 1793 (1992)
36. Ferry, D.K., Zhou, J.-R.: Form of the quantum potential for use in hydrodynamic equations for semiconductor device modeling. *Phys. Rev. B* **48**, 7944 (1993) The reader should be cautioned that the “DG term” in the Zhou-Ferry theory is correct only in 1D as may be seen by noting that its current density is related to that of DGC theory by $\mathbf{J}_{ZF} = \mathbf{J}_{DG} + \mu_n \hbar^2 [(\nabla^2 s) \nabla s - \nabla \nabla s \cdot \nabla s] / (3m_n q)$
37. Ancona, M.G., Bennett, B.R., Boos, J.B.: Scaling projections for Sb-based p-channel FETs. *Solid-State Electron.* **54**, 1349 (2010)
38. Radasavljevic, M., et al.: High-performance 40 nm gate length InSb p-channel compressively strained quantum well field effect transistors for low-power ($V_{CC} = 0.5$ V) logic applications. *IEDM Tech. Dig.*, 727 (2008)
39. Bennett, B.R., Ancona, M.G., Boos, J.B., Canedy, C.B., Khan, S.A.: Strained GaSb/AlAsSb quantum wells for p-channel field effect transistors. *J. Cryst. Growth* **311**, 47 (2008)
40. Watling, J.R., Brown, A.R., Asenov, A., Svizhenko, A., Anantram, M.P.: Simulation of direct source-to-drain tunneling using the density-gradient formalism: non-equilibrium Green’s function calibration. *Int. Conf. Sim. Sem. Proc. Devices (SISPAD)* 267 (2002)
41. Ancona, M.G.: Macroscopic description of quantum mechanical tunneling. *Phys. Rev. B* **42**, 1222 (1990)
42. Brown, A.R., Martinez, A., Seoane, N., Asenov, A.: Comparison of density gradient and NEGF for 3D simulation of a nanowire MOSFET. In: Proc. 2009 Spanish Conf. Elect. Dev. 140 (2009)
43. Hohn, T., Schenk, A., Wettstein, A., Fichtner, W.: On density-gradient modeling of tunneling through insulators. *IEICE Trans. Electron.* **E86C**, 379 (2003)
44. Ancona, M.G., Svizhenko, A.: Density-gradient theory of tunneling: physics and verification in one dimension. *J. Appl. Phys.* **104**, 073726 (2008)
45. Ancona, M.G.: Density-gradient analysis of field emission from metals. *Phys. Rev.* **46**, 4874 (1992)
46. Bender, C., Orszag, S.: *Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory*. Springer, New York (1999)
47. Ancona, M.G., Lilja, K.: Multi-dimensional tunneling in density-gradient theory. In: Proc. Workshop on Computational Electronics, vol. 38 (2005)
48. Ancona, M.G., Yergeau, D., Yu, Z., Biegel, B.A.: On ohmic boundary conditions for density-gradient theory. *J. Comput. Electron.* **1**, 103 (2002)
49. Wallmark, J.T., Marcus, S.M.: Maximum packing density and minimum size of semiconductor devices. In: Proc. Int’l. Electron. Devices Meeting, vol. 34 (1961)